VOL. XXII 1931

The Journal of Educational Psychology

Devoted Primarily to the Scientific Study of Problems of Learning and Teaching

---

BOARD OF EDITORS:

HAROLD RUGG, *Chairman*
Lincoln School of Teachers College
Teachers College, Columbia University

JAMES CARLETON BELL
College of the City of New York

FRANK NUGENT FREEMAN
University of Chicago

ARTHUR IRVING GATES
Teachers College, Columbia University

VIVIAN ALLEN CHARLES HENMON
Yale University

RUDOLF PINTNER
Teachers College, Columbia University

BEARDSLEY RUML
General Education Board

LEWIS MADISON TERMAN
Leland Stanford University

EDWARD LEE THORNDIKE
Teachers College, Columbia University

Published Monthly Except June to August by
WARWICK and YORK, Inc.,
Baltimore, Md.

TABLE OF CONTENTS

Association Factor in Intelligence Testing, The. S. Tolansky 321-333

Automatic Prediction of Scholastic Success by Using the Multiple Regression Technique with Electric Tabulating and Accounting Machines, The. David Segel 139-144

Certain Ambiguous Terms in Educational Psychology. Stephen Maxwell Corey 131-138

Comparison of White and Negro Children, A: Norms on Mirror-drawing for Negro Children by Age and Sex. R. J. Clinton 186-190

Conditioning of Overt Emotional Responses, The. Harold Ellis Jones. 127-130

Constant Changes in the Stanford-Binet IQ. Psyche Cattell 544-550

Correction, A 480

Development of Mental Ability at College-adult Level, The. Melvin B Wright 610-628

Effect of Order of Presentation on the Recall of Pictures. Daniel D. Droba 677-682

Effect of the Form of a Combination in the Learning of a Multiplication Table by Bright and Dull Children, The. F. T. Wilson. 536-537

Effect of the 6-22 44-22 6 Normal Curve System on Failures and Grade Values, The. J. De Witt Davis 636-640

Equivalence of Judgments to Test Items in the Sense of the Spearman-Brown Formula, The. H. H. Remmers. . 60-71

Errors, Difficulty, Resourcefulness, and Speed in the Learning of Bright and Dull Children. Frank T. Wilson . . 229-240

Experiment Designed to Test the Validity of a Rating Technique, An. Theodore Newcomb. 279-289

Experimental Comparison of the Study-test and Test-study Methods in Spelling, An. Arthur I. Gates . . 1-19

Experimental Study of the Influence of Motion Picture Films on Behavior, An. Frank N. Freeman and Carolyn Hoefer 411-425

Factors Involved in Children's Friendships. Gladys Gardner Jenkins 440-448

Factual Memory of Secondary School Pupils for a Short Article Which They Read a Single Time. Alfred G. Dietze and George Ellis Jones 580-598

Factual Memory of Secondary School Pupils for a Short Article Which They Read a Single Time. "Concluded." Alfred G. Dietze and George E Jones . . 667-676

Falsification of Age. A Factor in Child Guidance. Anna Cohen and Nathan Altrowitz . 470-478

Four Types of Examinations Compared and Evaluated. Alvin C. Eurich . . . 268-278

Further Data Concerning the Effect of Weighting Exercises in New-type Examinations. C. W. Odell . . . 700-704
Group Intelligence Test Suitable for Younger Deaf Children, A. R. Pintner 300-303
Growth of Social Perception in Different Racial Groups, The. W. N. Kellogg and B. M. Eagleson . . . 367-375
Handwriting of Indians, The. Thomas R. Garth . . . 705-709
Influence of an Audience upon Recall, The. Clara Burri 683-690
Influence of Blood Relationship and Common Environment on Measured Intelligence, The. Verner Martin Sims . . . 56-65
Influence of Jazz and Dirge Music upon Speed and Accuracy of Typing, The. Milton B Jensen . . . 458-462
Influence of the Assignment of Learning. David H Briggs and A. M. Jordan 659-666
Intelligence, Motivation, and Achievement. Austin H. Turney 426-434
Measurement of the Knowledge of Psychology before and after Formal Training, A. Calvin Hall 710-712
Mechanical "Aptitude" or Mechanical "Ability"? A Study in Method. O. L. Harvey 517-522
Method for Judging the Discrimination of Individual Questions on True-false Examinations, A. C. H. Whelden, Jr. and F. J. J. Davies 299-308
Mirror Reading as a Method of Analyzing Factors Involved in Word Perception. Miles A Tinker and Florence L. Goodenough. . . . 493-502
Modification of the Computation of the Multiple Correlation and Regression Coefficients by the Tolley and Ezekiel Method, A. Aaron Bakst 629-635
Nature of Intelligence, The J. H. Wilson . . 20-34
New Publications in Educational Psychology and Related Fields of Education. Conducted by Frances M. Foster
76-80, 157-160, 316-320, 389-393, 479-480
Note on Methods of Measuring Reliability, A. T. G Foran 384-387
Note on the Definition of the Harmonic Mean, A. Eugene Shen. 311-312
Note on the Standard Errors of the Standard Errors of Estimate and Measurement. Chester E. Kellogg and Kenneth W. Spence. 313-315
Objective Methods of Ranking Nursery School Children on Certain Aspects of Musical Capacity. Thomas F. Vance and Medora B. Grandprey 577-585
"Omission" as a Specific Determiner in the True-false Examination, The. C. C. Weidemann. . . 435-439
On Kin Resemblances in Physique vs. Intelligence. Herbert S. Conrad. 376-382

On Partial Correlation vs. Partial Regression for Obtaining the Multiple Regression Equations. HAROLD D. GRIFFIN 35-44
✓One More Study of Permanence of Interest. HARVEY C. LEHMAN and PAUL A. WITTY 481-492
Organismic Psychology and Educational Theory. KENNETH SELTSAM 351-359
Our Need of Some Science in Place of the Word "Intelligence." C. SPEARMAN. 401-410
Parental Age and Intelligence of Offspring. MINNIE LOUISE STECKEL. 212-220
Practice versus Grammar in the Learning of Correct English Usage. PERCIVAL M. SYMONDS 81-95
Prognosis of Abilities to Solve Exercises in Geometry WINONA M. PERRY. 604-609
Proof That the Point from Which the Sum of the Absolute Deviations Is a Minimum Is the Median, A. PAUL HORST 463-464
Publications Received 394-400
Relation between Use of Different Parts of Speech in Written Composition and Mental Ability. CHARLES P. LOOMIS and ANNA MAY MORAN. 465-475
Relative Effort of Children of Native vs. Foreign Born Parents, The S. EDSON HAVEN 523-535
Relative Influence of Visual and Auditory Factors in Spelling Ability, The. GEORGE W. HARTMANN. 601-609
Reliability of Integration Index Differences JOHN W. DICKEY 209-211
Reply to Professor Kelley. KARL J. HOLZINGER. . . 455-457
Reply to Some Recent Criticisms by Professor Spearman, A. H. D. CARTER 118-119
Retention of Mirror-reading Ability after Two Years, The. FLORENCE L. GOODENOUGH and MILES A. TINKER. . . 503-504
Routine Computation of Partial and Multiple Correlation, The. RAYMOND FRANZEN and MAHEW DERRYBERRY 641-651
Sampling Error of Tetrad Differences. C. SPEARMAN . 388
Scale of Militarism-pacifism, A D. D DROBA . . . 96-111
School Achievement in Relation to Mental Age—A Comparative Study. ANDREW W. BROWN and CHRISTINE LIND. . . . 561-570
Self-cultivation and the Creative Act: Issues and Criteria. HAROLD RUGG . . . 241-254
Semi-logarithmic versus Linear Plotting of Learning Curves. RICHARD W. HUSBAND 72-75
Sex Differences: Collecting Interests. PAUL A. WITTY and HARVEY C. LEHMAN 221-228
Shrinkage of the Coefficient of Multiple Correlation, The. SELMER C. LARSON. . . . 45-55

Sigmas of Combined Distributions Calculated from Sigmas, Means, and Frequencies of Component Distributions, The. C. R. Garvey . . . 307–310
Significance of a Difference between "Matched" Groups, The. E. F. Lindquist . . . 197–204
Some Relationships between Algebra and Geometry. Dorris May Lee and J. Murray Lee . . . 551–560
Standard Error of the Means of "Matched" Samples, The. Samuel S. Wilks . . . 205–208
Studies in Handedness: III. Relation of Handedness to Speech. R. H. Ojemann . . . 120–126
Study Habits of Teachers College Students. Hugh M. Bell . . . 538–543
Study in Reversing the Handedness of Some Left-handed Writers, A. Norma V. Scheidemann and Hazel Colyer. 191–196
Study of Classroom Behavior, A. Willard C. Olson . . . 449–454
Study of Questionnaire Technique, A. Frank K. Shuttleworth . . . 652–658
Technique for Experimentation on Guessing in Objective Tests, A. Louis Granich . . . 145–156
Tetrad-differences for Non-verbal Subtests. William Stephenson . . . 167–185
Tetrad-differences for Verbal Subtests. William Stephenson . . . 255–267
Tetrad-differences for Verbal Subtests Relative to Non-verbal Subtests. William Stephenson . . . 334–350
Thorndike's C.A.V.D. is Full of G. Karl J. Holzinger 161–166
Vocabulary Study of Biology Notebooks of Fifty Representative Secondary Schools in New York State, A. Don O. Baird . . . 512–516
Vocational Interests and Types of Ability. Ralph H. Gundlach and Elizabeth Gerum . . . 505–511
What is Meant by a G Factor? Truman L. Kelley 364–366
What the Theory of Factors Is Not. C. Spearman 112–117
Why Otis' "IQ" Cannot be Equivalent to the Stanford-Binet IQ. Psyche Cattell . . . 599–603

BOOK REVIEWS

Bolton, Frederick E., *Adolescent Education*. (Dennis C. Troth). 470
Buros, F. C. and O. K., *Expressing Educational Measures as Percentile Ranks*. (Laurance F. Shaffer) . . . 158
Gast, I. M. and Skinner, H. C., *Fundamentals of Educational Psychology*. (Gertrude Hildreth) . . . 70

Gesell, Arnold, *The Guidance of Mental Growth in Infant and Child.* (Gertrude Hildreth) 391
Gifford and Shorts, *Problems in Educational Psychology.* (P. Sandiford) 316
Hartshorne, Hugh, May, Mark A. and Shuttleworth, Frank K., *Studies in the Organization of Character.* (Donald Snedden). 77
Hollingworth, H. L., *Abnormal Psychology, Its Concepts and Theories.* (H. Meltzer). 389
Hurd, Archer Willis, *Problems of Science Teaching at the College Level.* (Leonard B. Wheat). 319
Inskeep, Annie Dolman, *Child Adjustment.* (Gertrude Hildreth) 159
Monroe, DeVoss and Reagan, *Educational Psychology.* (P. Sandiford) 316
Paterson, D. G., Elliott, R. M., Anderson, L. D., Toops, H. A., and Heidbreder, E., *Minnesota Mechanical Ability Tests.* (O. L. Harvey) 317
Runnels, Ross O., *Manual for Determining the Equivalence of Mental Ages Obtained from Group Intelligence Tests.* (Laurance F. Shaffer) 158
Waples, Douglas and Tyler, Ralph W., *Research Methods and Teachers' Problems.* (Gertrude Hildreth). 79
Wheeler, Raymond Holder, *The Science of Psychology.* (John N. Washburne). 157

INDEX TO AUTHORS

Altrowitz, Nathan . . . 476
Baird, Don O . . . 512
Bakst, Aaron . . . 629
Bell, Hugh M . . . 538
Briggs, David H . . . 659
Brown, Andrew W . . . 561
Burri, Clara . . . 653
Carter, H. D . . . 118
Cattell, Psyche . . . 541, 599
Clinton, R. J . . . 180
Cohen, Anna . . . 470
Colyer, Hazel . . . 101
Conrad, Herbert S . . . 376
Corey, Stephen Maxwell . . . 131
Davies, F. J. J . . . 200
Davis, J. De Witt . . . 636
Derryberry, Mahew . . . 641
Dickey, John W . . . 209
Dietze, Alfred G . . . 580, 667
Droba, Daniel D . . . 96, 677
Eagleson, B. M . . . 307
Eurich, Alvin C . . . 288
Foran, T. G . . . 383
Franzen, Raymond . . . 641
Freeman, Frank N . . . 411
Garth, Thomas R . . . 705
Garvey, C. R . . . 307
Gates, Arthur I . . . 1
Gerum, Elizabeth . . . 505
Goodenough, Florence . . . 493, 603
Grandprey, Medora B . . . 577
Granich, Louis . . . 146
Griffin, Harold D . . . 35
Gundlach, Ralph H . . . 505
Hall, Calvin . . . 710
Hartmann, George W . . . 601
Harvey, O. L. . . . 517
Haven, S. Edson . . . 523
Hoefer, Carolyn . . . 411
Holzinger, Karl J . . . 101, 455
Horst, Paul . . . 403
Husband, Richard W . . . 72
Jenkins, Gladys Gardner . . . 440
Jensen, Milton B . . . 458
Jones, George Ellis . . . 580, 667
Jones, Harold Ellis . . . 127
Jordan, A. M . . . 659
Kelley, Truman L . . . 364
Kellogg, Chester E . . . 313
Kellogg, W. N . . . 307

Larson, Selmer C 45
Lee, Dorris May 551
Lee, J. Murray 551
Lehman, Harvey C 221, 481
Lind, Christine 361
Lindquist, E F 197
Loomis, Charles P 465
Moran, Anna May 465
Newcomb, Theodore 279
Odell, C W 700
Ojemann, R. H 120
Olson, Willard C 449
Perry, Winona M 604
Pintner, R. 360
Remmers, H. H 66
Rugg, Harold 241
Scheidemann, Norma V 191
Segel, David 139
Seltram, Kenneth 361
Shen, Eugene 311
Shuttleworth, Frank K 652
Sims, Verner Martin 56
Spearman, C 112, 388, 401
Spence, Kenneth W 312
Steckel, Minnie Louise 212
Stephenson, William 167, 255, 334
Symonds, Percival M 81
Tinker, Miles A 163, 503
Tolansky, S 321
Turney, Austin H 426
Vance, Thomas F 577
Weidemann, C C 435
Whelden, C H, Jr 290
Wilks, Samuel S 305
Wilson, Frank T 224, 636
Wilson, J H 30
Witty, Paul A 221, 481
Wright, Melvin B 610

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII January, 1931 Number 1

AN EXPERIMENTAL COMPARISON OF THE STUDY-TEST AND TEST-STUDY METHODS IN SPELLING

ARTHUR I. GATES

Teachers College, Columbia University

The present study is one of a series of investigations planned to inquire into the possibilities of improving instruction in spelling. Investigations accepted as dissertations for the Doctorate at Teachers College by Robert Thompson, Herbert A. Carroll, James E. Mendenhall, Claire T. Zyve, Ina C. Sartorius, and W. H. Coleman,[1] deal with closely related problems, and others will be completed shortly.

The present investigation is an inquiry concerning the gross efficiency of the two most widely used general methods: the Pre-study or Study-test method and the Pre-test or Test-study plan. In some measure, the investigation provides data which suggest the inherent limitations and merits of the two rival procedures and which indicate possible improvements in each. The study also provides several clues for further research for the purpose of appraising modifications of possible value.

The literature affords few experimental comparisons of the Pre-study and the Pre-test methods. Keener's study[2] based upon nine-hundred seventy-six pupils in Grades II to VIII of the Chicago schools seems to show that there is a slight superiority of the Pre-study method in Grades II and III, and superiority of the Pre-test plan in Grades IV to VIII. A study by Woody,[1] confined to pupils in Grades VI, VII, and VIII gave no consistent advantage to either plan. Kilzer[2] in a study confined to ninth-grade pupils found, according to his methods of computing results, that while the test-study method gave better immediate results, the advantage was not apparent in tests after an interval of six months.

---

[1] Thompson, R.: "The Effectiveness of Modern Spelling Instruction," Teachers College, Columbia University, Contributions to Education, No. 436; Carroll, H. A.: "Generalization of Bright and Dull Children: A Comparative Study with Special Reference to Spelling," Teachers College, Columbia University, Contributions to Education, No. 439; Mendenhall, J. E.: "An Analysis of Spelling Errors," Teachers College Bureau of Publications, 1930; Zyve, C. T.: "A Comparative Study of Methods" (in press); Sartorius, I. C.: "The Bases of Generalization in Spelling" (in preparation); Coleman, W. H.: "Studies of Spelling Vocabularies" (in preparation).

[2] Individual Method vs. Group Method of Teaching Spelling. *Fourth Year Book of the Department of Superintendence*, Washington, D. C., 1926.

THE STUDY-TEST PROGRAM

For the Study-test procedure the same words were used as for the Test-study method. The number per week varied with the grade. Each week's assignment was divided into four equal parts. One of these short lists was studied on Monday, Tuesday, Wednesday, and Thursday, respectively. Friday was devoted to a test on the week's assignment, followed by review when time permitted.

The method of introducing the words was substantially as follows:

1. The teacher pronounced the word clearly. If the word contained two or more syllables, it was pronounced by syllables.
2. The word was used in one or more sentences. In some cases pupils were asked to use the word in a sentence.
3. The teacher wrote the word on the blackboard and had the children say it.
4. The pupils looked at the word and said it, syllable by syllable, to themselves.
5. The pupils looked at the word and said the letters to themselves. The pupils were encouraged to group the letters by syllables as they said the letters.
6. They looked at the word as a whole as they said it to themselves.
7. Same as (5).
8. Pupils closed their eyes and said the letters to themselves. In some instances the teacher asked one pupil to do this aloud while the others listened.
9. Pupils wrote the word as they said the letters to themselves. They proceeded by syllables whenever possible.
10. They compared their written word with the correct form on the board.
11. Pupils covered their word, wrote it again and compared it with the correct form.
12. They repeated (11) until they could write the word without error.

---

[1] The Evaluation of Two Methods of Teaching Spelling. *Fifteenth Yearbook of the National Society of College Teachers of Education*, University of Chicago Press, 1927.

[2] Kilzer, L. R.: The Test-study *vs.* the Study-test Method in Teaching Spelling. *School Review*, 1926, pp. 521–525.

The above plan was modified in details to meet the particular needs of the different groups, which varied from the dullest second-grade to the brightest eighth-grade classes.

Near the end of the period the pupils were given a test of the words studied during the day. If a word was missed by more than half the class, it was added to the list for the next day.

On Friday a test of all words taught during the week was conducted. Words frequently missed were put into a review list. Those proving to be particularly difficult were retaught, if possible, during the following week. Others were retained for a review period about a month later.

THE TEST-STUDY PLAN

The Test-study plan employed was in general outline the plan recommended by Horn[1] in 1919 and since widely used. The schedule is as follows:

Monday.—Test on all words in the new assignment for the week.

Tuesday.—Individuals study the words missed on the Monday test. Pupils making no errors on the Monday test are excused.

Wednesday.—Test on new and review words All pupils take this test

Thursday.—Study of words missed on Wednesday test. Pupils making no errors are excused

Friday—Test on same words as used on Wednesday for all pupils. Study missed words as far as time permits

Once a month a review test was given comprising the most difficult words used during the preceding month.

THE EXPERIMENTAL SCHEDULE

The experiment was conducted during the period from February 1 to June 30, 1928. The period was divided into two parts. Each period consisted approximately of nine weeks for teaching and one week for testing. During the first period half of the classes used the Study-test method and half the Test-study. During the second period the methods were reversed. Thus, each pupil spent approximately ten weeks with one method and ten with the other. This scheme makes two types of comparisons possible: (1) Results obtained from groups using the two different methods at the same time and with the same words may be compared, and (2) results from the same group obtained from one method in the first period may be compared with those obtained under the other method in the second period.

---

[1] Principles of Method in Teaching Spelling as Derived from Scientific Investigations *Eighteenth Yearbook of the National Society for the Study of Education*, Part II

THE SUBJECTS

The subjects comprised originally nearly 1800 pupils in fifty-four classes in Grades II to VIII of Public School 210, Brooklyn, under Principal Frederick C. Graham. The number of records fully complete and usable in computing results were as follows:

First period	49 classes	1558 pupils
Second period	49 classes	1678 pupils
Total number of complete records		3236

For the equivalent group studies (which necessitated the elimination of records to secure equivalent scores) the total number of complete pupil-records was 2900.

PRELIMINARY AND FINAL TESTS

The same list of words was used for a preliminary and final test. The list consisted of fifty words chosen at random from the words to be studied by each class during the ten weeks of the experiment. The tests were given by the familiar column dictation method. The number of words correct was multiplied by two to convert the scores into percentages correct.

The IQ's obtained by group-testing in the routine work of the school were secured from the school records and used in comparing groups. That the initial spelling scores for the classes correspond very closely to the intelligence ratings of the classes at a given grade level is apparent in the data of Table I. It is also obvious that, probably as a result of frequent examinations with the group tests, the IQ's, beginning with Grade III, run high. It is believed, however, that the population of the school as a whole is close to average in native intellectual capacity.

RESULTS

Comparison of Classes within a Given Grade. Since the pupils in the school are classed on the basis of intelligence tests, it was impossible to secure groups of unselected pupils within a given grade. A schedule was therefore adopted which would give as nearly as possible two equivalent combinations of classes within each grade. This was

TABLE I.—COMPARISONS OF CLASSES USING STUDY-TEST AND TEST-STUDY PLANS

Study-test plan					Test-study plan				
Number	IQ	Initial	Final	Gain	Number	IQ	Initial	Final	Gain
Grade II									
32	94	55	97	44.5	42	88	34.4	85.6	41.2
26	80	12.5	58.5	46.0	28	84	22.1	63.8	41.7
*58	87.2	34.0	79.9	45.9	*70	86.6	29.7	75.0	45.3
Low Grade III									
38	109.4	52.8	90.5	37.7	39	121.3	73.4	95.1	21.7
32	99.8	31.2	73.5	42.3	27	91.3	28.4	66.1	37.7
*70	105.0	44.4	83.1	38.7	*66	108.5	52.5	85.6	33.1
High Grade III									
35	126.1	54.5	68.3	13.8	34	106.6	64.4	93.6	29.2
36	110.6	69.4	92.3	22.9	27	102.0	49.3	79.1	29.8
22	91.5	42.9	73.0	30.1					
*93	112.2	64.4	90.2	25.8	*61	105.2	56.5	87.1	27.9
Grade IV									
32	118.6	60.3	97.2		24	96.3	67.5	91.9	24.0
24	92.3	53.3	89.2	35.0	21	83.6	14.9	82.1	27.5
31	107.8	64.0	88.2		44	121.2	80.5	99.1	
28	86.7	47.6	89.7		19	98.5	63.5	94.5	
*115	101.35	61.8	90.8		*108	101.15	64.05	91.2	
Grade V									
46	125.3	70.8	97.8		41	110.8	77.1	94.2	
30	102.1	58.0	93.6		28	92.0	61.5	88.0	
20	101.1	76.3	95.8		27	87.5	58.7	84.5	
*96	110.1	72.1	96.1		*96	100.4	61.2	91.1	
Grade VI									
30	125.7	76.3	99.2		28	118.6	88.3	98.0	
30	104.2	59.0	88.5		42	115.1	66.1	90.7	
27	108.1	65.8	91.0		31	90.4	55.0	70.7	
					37	106.6	61.5	94.1	
*87	112.7	67.2	92.1		*138	108.1	68.7	92.05	
Grade VII									
30	120.1	81.0	94.8		37	121.6	72.1	91.8	
35	109.3	68.1	92.0		23	102.0	61.7	87.5	
30	130.4	80.4	97.7		32	120.0	70.0	94.2	
28	98.1	68.1	91.3		25	108.0	65.7	89.2	
*118	122.0	76.05	94.95		*117	114.5	70.1	92.1	
Grade VIII									
34	137.5	74.8	94.5		22	99.5	61.8	89.5	
31	107.2	60.3	90.9		30	115.2	70.4	94.0	
30	99.2	56.8	89.0		37	120.8	73.0	96.1	
31	117.5	62.0	94.1		28	110.6	57.0	94.9	
*128	115.1	64.8	94.0		*117	111.2	65.5	94.6	

---

* Lines marked with asterisk are totals or averages

usually done by combining the brightest and dullest classes in one group and the middle classes in the other. The extremes were assigned to the Study-test plan and the middle groups to the Test-study plan in one grade, the reverse in the next, and so on alternately. Defaults in some of the classes upset the scheme in certain classes.

The records available for the first experiment are shown in Table I. In this table are given for each group the number of pupils, the mean IQ, the mean initial spelling score in percentage correct, the mean final spelling score, and the mean gain. The gain is merely the difference between the mean initial and mean final scores. The last line under each grade gives the number of pupils, the mean IQ, initial and final scores, and gain for all the pupils of that grade. These figures were obtained not from the preceding class-scores but from the individual records of the pupils in the groups.

Since a study by Robert S. Thompson[1] revealed serious difficulty in comparing gains of groups which differ in initial scores, the data in Table I are not readily interpreted when the initial scores are unequal. Dr. Thompson found that, other things being equal, pupils obtaining a low initial score gain more, in terms of percentages of correct spellings, than those obtaining higher initial scores. He found that, in general, equal teaching and learning would produce equal advances in the interval between the initial percentage correct and one hundred percentage correct. Thus, if one group advances, say, half of the distance from initial score forty percentage to one-hundred percentage correct or to seventy percentage correct, another group with an initial score of sixty should advance to $60 + \left(\frac{100 - 60}{2}\right)$ or 80 as the result of equal learning. While these relations given by Thompson could be applied in this case, it was felt that where differences are as small as these appear to be the more laborious process of sifting the pupils down to groups equivalent in initial scores would be more reliable. Table I is offered for those who wish to make comparisons of classes by utilizing Thompson's data or other means.

Comparison of Gains Made by Groups of Equivalent Initial Ability.—In Table II the results of both studies are shown. In each case pupils were first arranged by grades. Then pupils following the Study-test plan were matched in initial scores with pupils following the Test-study plan. This usually resulted in a surplus of initial scores in one group and low scores in the other, which had to be eliminated. The plan pursued was not an exact one-per-one matching, but one which gave the largest number of cases without appreciably disturbing the equivalence of the groups in mean and SD of initial scores. Thus for example, one group might be given three pupils with scores of thirty-nine, forty, forty-one respectively as equivalent to two scores of forty, when the effects of this assignment could be counterbalanced. Consequently, although the groups are substantially equivalent, the number of pupils in the group are not always the same

---

[1] Thompson, R.: "The Effectiveness of Modern Spelling Instruction." Teachers College, Columbia University Contributions to Education, No. 436.

The results revealed by Table II may be summarized by the statement that in so far as differences are indicated at all, the Study-test method produces larger gains in Grade II and low Grade III and the Test-study plan yields greater gains from high Grade III to the Grade VIII inclusive.

The advantages of the Study-test plan in the second and lower third grades (*i.e*, first half of the third grade) are neither large nor, in terms of the Standard Error, highly reliable. Since the groups are evenly matched in initial spelling scores and since the advantage in IQ's, if any, is enjoyed by the Test-study groups, and since the same teachers taught the groups under both plans, the consistency of the superiority in gains shown by the Study-test method is indicative of a genuine, even if small, difference. The average superiority of the Study-test plan in the two grade levels based on four hundred and seventy-seven pupil-records is 1.95, or approximately two per cent. The advantage is slightly higher (2.11) in the second grade and lower (1.36) in Grade III. Whether this small advantage, assuming it to be genuine, is sufficient to demonstrate superiority of the Study-test plan in general in these grades is a topic for consideration later.

Beginning with the second half of Grade III and continuing to the end of Grade VIII, the Test-study plan shows slightly greater gains. Aside from one small setback in Grade IV and another in Grade VII, the Pre-test scheme shows to advantage in all of the dozen comparisons. The surplus in favor of this plan is consistently small and often lacks "satisfactory" statistical reliability. The advantages, beginning with the lowest grades are: 0 53, 1.45, −0.50, 2.00, 2 80, 1.91, 2.84, 3.57, 1.76, −0 80, 0.41, 1.11. The average of these figures, based on 2423 pupil-records, is 1.40. The series shows no consistent deviation from this average surplus in favor of the Test-study plan. The plan, in other words, beginning at the middle of Grade III, seems to work as well, relatively, in the lower as in the upper grades.

SUMMARY OF TABLE II.—DIFFERENCES IN FAVOR OF EACH METHOD

Grade	Number of cases	IQ lower S-T	IQ [illegible] T-S	[illegible] S-T	[illegible] T-S
II	118		0.74	2.53	
II	117	0.22		0.27	
IIIA	120	1.10		[illegible]	
IIIA	147		1.70	1.18	
IIIB	137		0.50		.83
IIIB	141	1.00			1.45
IV	213	.40		.50	
IV	214	1.63			2.00
V	177		2.00		2.20
V	183	.40			1.91
VI	210		1.00		2.43
VI	216	1.20			3.57
VII	220		.70		1.76
VII	240		1.00	.90	
VIII	237	3.20			.44
VIII	228		1.40		1.10
Total	2900	9.72	8.54	8.23	18.09

COMPARISON OF RESULTS IN BRIGHT AND DULL CLASSES

An inspection of Table I suggests the possibility that the Study-test plan shows to greatest advantage in classes of duller pupils, whereas the Test-study method is more suited to the bright. Table III is obtained by taking from each grade level the brightest and the dullest classes. Each class was taught by the same teacher first by one method and then by the other. Since the initial tests were composed of different words and given at different times, pupils' initial scores were often different. To secure groups as nearly equivalent as possible on the initial test scores the records were sifted as before.

Table III shows that in Grades II, III, and IV, the dullest pupils made greater gains when taught by the Study-test plan. In Grades V, VI, VII, and VIII they do as well when taught by the Test-study method. The brightest classes on the whole make larger gains when working by the Test-study plan. There are two classes, in Grade II and Grade VI, in which Test-study shows no advantage. The average superiority of the method is 1.61, a figure slightly greater than the average gain of 1.40 percentage for the entire population of the school.

In terms of the gains in ability to spell the words after an interval of from one to ten weeks[1] after the words were studied the results show the Study-test method to be slightly better in the second and first half of Grade III, whereas the Test-study method is superior thereafter. In the case of the dullest of the four or five classes at the high third and fourth grade levels the Study-test plan is slightly superior, whereas in the case of the brightest of these groups, the Test-study plan is equal to Study-test in the second grade and slightly superior thereafter. Were no other factors than these results taken into account, the recommendation would be to use the pre-test plan for bright pupils from the beginning, for average pupils from the middle of Grade III, and for the slowest pupils from the beginning of Grade V and to use the Pre-study method in the remaining classes. Since the advantage of either method is small in terms of specific spelling gains, it will be sensible to consider certain other factors.

Let us consider first certain limitations urged against the Pre-test method in comparison with the Pre-study procedure. One frequently mentioned limitation of the Test-study method lies in the fact that a single dictation test is not a wholly reliable means of determining which words a child can spell correctly. In the first place the test puts the child into a state of concentration on spelling. A word which he might misspell in casual writing or even in a dictation exercise in which attention is partly devoted to meaning or composition, or both, may be correctly spelled when attention is entirely devoted to spelling. Thus words insufficiently mastered are omitted from study. In reply to this objection it may be said, however, that the Test-study plan provides three tests (on Monday, Wednesday, and Friday) and that words known well enough to be spelled correctly three times probably need no further systematic drill

It is urged similarly that children who are not certain as to which of two or three alternatives to use will succeed frequently in getting the right form by chance Thus, Kilzer[2] found that of words misspelled in a second test, thirty-one per cent were spelled correctly on the first test. Nearly a third of the words misspelled on the first test apparently were not perfectly known and, it is urged by some, should have been studied. In this connection the plan of giving the pupils a brief preview of the words, as by reading them in printed

---

[1] Since the final test included an equal number of words from each of the weekly lists, it occurred at intervals of from one to ten weeks after the words were studied.

[2] *School Review*, 1926, pp 521 525

TABLE III.—COMPARISON OF GAIN MADE BY BRIGHT AND DULL PUPILS

Method	Dull classes						Bright classes					
	No.	IQ	Initial score	Final score	Gain	SD gain	No.	IQ	Initial score	Final score	Gain	SD gain
Grade II												
S-T	26	80	12.5	58.5	46.0	20.4	32	94	55.0	97.0	42.0	18.5
T-S	26	81	16.4	57.4	41.0	21.8	30	95	56.2	97.8	41.6	19.4
Diff (1-2)					5.0						.4	
Diff./SD Diff.					.88						.05	
Low Grade III												
S-T	30	90.2	30.2	70.5	40.3	21.3	36	123.1	75.0	96.0	21.0	16.7
T-S	27	91.3	28.4	66.1	37.7	20.1	39	121.3	73.4	98.1	24.7	17.4
Diff (1-2)					2.6						−3.7	
Diff./SD Diff.					.47						.94	
High Grade III												
S-T	22	91.5	42.9	73.0	30.1	18.3	35	126.1	84.5	98.3	13.8	14.9
T-S	27	89.5	44.5	72.1	27.6	17.1	33	125.7	80.4	97.6	17.2	15.3
Diff					2.5						−3.4	
Diff./SD Diff.					.49						.92	
Grade IV												
S-T	25	83.4	42.8	73.9	31.1	17.4	35	123.0	80.5	97.4	16.9	12.4
T-S	24	83.6	44.9	72.4	27.5	14.8	33	123.2	80.5	99.1	18.6	14.7
Diff					3.6						−1.7	
Diff./SD Diff.					.79						.31	

Grade V

S-T	31	88 0	56 3	86 5	30 0	15 9	38	125 3	79 8	97 8	18.0	18 3
T-S	27	87 5	58 7	88 5	29 8	16.4	34	123 1	79 5	98 6	19 1	14 8
Diff (1-2)					2						−1 1	
Diff / SD Diff					05						25	

Grade VI

S-T	33	90 5	55 2	78 5	23 3	15 9	30	148 2	87 1	98 2	11 1	14 4
T-S	31	90 3	55 6	79 7	24 1	12 6	28	148 6	88 3	95 6	10 3	12 8
Diff (1-2)					− 8						8	
Diff. / SD Diff					22						22	

Grade VII

S-T	28	98 4	68 1	91 3	23 2	18 2	39	126 1	81 6	94 8	13 2	15 3
T-S	33	98 0	67 3	90 4	23 1	13 4	35	125 7	80 7	95 4	14 7	14 9
Diff (1-2)					1						−1 5	
Diff / SD Diff					03						44	

Grade VIII

S-T.	30	99 2	56 8	89 9	33 1	17 2	33	137 5	73 8	94 5	20 7	16 4
T-S	32	97 4	54 3	88 3	34 0	18 1	35	138 0	75 2	98 2	23 0	15 4
Diff (1-2)					− 9						−2 7	
Diff. / SD Diff					21						73	

columns or in context, has been attacked, inasmuch as it would [illegible] the number of words which a pupil could spell immediately without thorough mastery. The writer, in a small test of the matter, found that twenty-one per cent of the words spelled incorrectly on a second test were correct on the first test without preview, whereas the percentage was thirty after the children were given two seconds per word for a preview of the list. Kilzer's finding that eighty-three percentage of the words misspelled after an interval of six months were correctly spelled in a test immediately after a preview has been interpreted similarly.

These would be strong arguments were the study for the week determined entirely by the single Monday test. The additional tests on Wednesday and Friday seem to be excellently adapted to take care of the conditions here revealed. The plan seems fairly well to adjust the amount of study to the needs. With rare exceptions the words missed on the first test will be most poorly known and hence presumably most in need of the greatest amount of study on Tuesday and thereafter in case of a repeated failure on Wednesday. Contrariwise, the mere fact that a child spells a word correctly on Monday, even if he fails later, is evidence of a greater degree of learning than if he failed entirely. A word spelled correctly on both Monday and Wednesday is, with rare exceptions, still better known, and a word spelled correctly on all three days is probably known as well as words should ever be taught through direct drill in the spelling lesson. If such a word is misspelled six months later, the explanation is probably to be found in the interfering influence of words subsequently learned or to lack of use outside of the spelling lessons during the period. In the latter case the remedy is not more drill but better selection and placement of words. A word not used until six months after it has been taught has been introduced at least six months too soon. Instead of studying it more, the pupils should not have studied it at all at that time.

With reference to the management of the easy words—that is, words already known well enough to be spelled correctly in one or more of the three tests—the Test-study plan seems to the writer to have all the better of the argument. It is, in fact, nicely designed to save the pupil from needless overlearning of words he can already spell with some measure of consistency. The three tests will rarely let a word slip through without practice if it really needs drill at the time. No method can control all of the exigencies of future events.

Concerning the treatment provided for the most poorly known words, the advantage of the Test-study plan is usually admitted. This plan shows on Monday, with rare exceptions, the words that a pupil is least able to spell from the whole week's list and enables a pupil to concentrate on these as much as needed up to the limit of the full week. By saving time which the Study-test plan devotes to study of the easier words, this plan would seem to enable the pupil to direct his study to the words most in need of it.

At the same time the Test-study plan tends to exaggerate to the limit the effects of misplacement of words. The grade placement of words is far from a satisfactory condition, and in most schools many words are doubtless taught long before they are actually needed in writing. Such words will mainly be difficult words to spell because they have been used but seldom or never. The Test-study plan requires the pupil to devote a maximum amount of time in study of these words. Words which will not be used for a long time will now either be forgotten before they are needed or overlearned greatly to maintain their life until they are called for. In either case the time spent in drill is largely wasted.

The Study-test plan, then, may tend to waste time on the study of "easy" words already known well enough or nearly well enough for successful usage, whereas the Test-study plan tends to waste time upon study of "hard" words which ought not to be studied until later. Evidence in support of this view may be found in Thompson's study[1] of the causes of differences in the spelling difficulty of words. The present writer believes that his results may best be interpreted as indicating that "easy" words, on the whole, are those most frequently used in writing, and "hard" words those least frequently used prior to a given test.

With the limitation in grade placement now in effect, it is probable that one of these deficiencies is about as bad as the other. Theoretically, then, the contest may be called a draw up to this point.

It should be pointed out, however, that the difficulty with the Test-study plan is not as intrinsic and irremedial as the limitation of the Study-test method. Remove improper grade placement, and this wastefulness of the Test-study plan largely disappears. This will not be so fully true of the Study-test plan. Despite the probability that better grade placement will tend to produce daily or weekly lists of words of more uniform difficulty, on the average the words will still be of unequal difficulty to individuals in the class. It is precisely in the function of adjustment to individual differences in mastery of the various words in an assignment that the superiority of the Test-study method lies. The most perfect grading system will scarcely make all children equally good spellers or a given child able to spell equally well (or poorly) all of the words in a given assignment. With a perfect grading system, the Test-study plan would find a more ample defense than it does under present mediocre grade placement of spelling words.

---

[1] *Op. cit*

Another defect attributed to the Pre-test plan more than to the Pre-study plan is the loss of study upon words of all degrees of difficulty due to failure of the pupil or teacher, or both, to detect errors in the test results. Kilzer pointed out the frequency of such errors when the pupils correct one another's papers. They are likely to appear also when the teacher corrects the spelling. It may again be said, however, that while errors are likely to go unnoticed occasionally in one test, it is not very likely to be missed three times in succession. The Test-study plan makes rather adequate provisions for catching such mistakes in time to give a word at least one day of practice. The plan of having pupils correct their own spelling by comparing their test papers with the printed text has several features, moreover, of educational value. Apparently what is needed is to encourage the pupil to develop a higher standard of accuracy in scoring his spellings. Such a habit would be useful through life as an aid in the improvement of ability to spell.

Another objection offered to the Pre-test is that it makes initial errors not only a necessary but a natural result. Since the spelling of the words is not always known, errors are inevitable. Since the making of errors is a necessary result—taken for granted—it seems also natural. Thus ill-effects are alleged to be the issue. The first is the inculcating of an attitude of tolerance toward mistakes in spelling. Pupils are induced to look upon their mistakes with complacency. The second ill-effect lies in the fact that the Pre-test forces the pupil to "practice errors" In this connection the alleged tendency of initial errors to persist is often emphasized. The Study-test plan, on the other hand, by teaching before testing, tends to set up standards of accuracy, to prevent initial errors, and to prevent the practicing of errors.

While the writer believes that the importance of these tendencies of the strict Pre-test method has often been overestimated, he is inclined to believe that they are genuine and, in some measure, potent. To what extent the making of errors during the Pre-test tends to produce a tolerance for misspellings, if at all, is difficult to show. It is not impossible, indeed, that testing, by revealing errors, tends to cultivate an opposite inclination to correct them and to develop a sensitivity to errors which ordinary writing does not properly cultivate. The general tendency for initial errors to persist is apparently far less strong in spelling than in some other functions. Errors in spelling tend to be variable rather than constant. If a child cannot spell a word he most frequently spells the word phonetically. Most words may be spelled phonetically in almost innumerable ways, as Horn[1] and others have shown, and they are spelled in many of these ways in repeated attempts by the same pupil. As Woody found, moreover,[2] the errors made by children in preliminary tests do not tend appreciably to persist in their original form.

Another consideration is the relation of the two methods to the interest and effort of the teacher and pupils. Keener's finding that "the majority of teachers . . . were very markedly in favor of the [Test-study] method" and that "they favored it because of the greater interest on the part of the pupils . . . and opportunity of giving help where it was needed" is borne out in Public School 210. Although the Study-test plan was more of a novelty to the pupils, they favored the Test-study plan by nearly ten to one in a sample of votes taken.

A problem of great importance is that of the amount of time consumed by the two methods. Adherents of the Study-test method point out that relatively more classroom time is available for *study* when this method is used because less time is spent in testing. As commonly used, the Study-test plan requires two tests per word per week, whereas the Test-study plan requires three weekly tests

Adherents of the Test-study plan reply, however, that their plan requires less time in absolute terms—fewer pupil-minutes of work per week. The reason for this is that pupils are not required to study all words, as in the Pre-study plan, but only those missed on a test. Thus, if a pupil misses no words on the Monday test he is excused entirely from the Tuesday study period. All of those successful on Wednesday are excused from the Thursday period. Keener found in his study that about twelve per cent were excused from both periods.

---

[1] A Source of Confusion in Spelling. *Journal of Educational Research*, Jan., 1929, pp. 47–50

[2] The Evaluation of Two Methods of Teaching Spelling. *Year Book of the National Society of College Teachers of Education*, 1926

In the classes sampled in the present study about ten per cent were excused from both periods and about eighteen per cent from the second period.

The saving of pupils' efforts is by no means fully represented by the number of children who know how to spell all of the words in a week's assignment in advance of instruction. Nearly all of the pupils know some of the words. In Table II the scores on the initial test represent the average per cent of words to be taught during a ten-week period which are known in advance of study. These represent the percentage which they would not need to study under the Pre-study plan. They are approximately as follows: Grade II, thirty-five per cent; Grade III, fifty-five per cent; Grade IV, sixty-five per cent; Grade V, sixty-nine per cent; Grade VI, seventy-one per cent; Grade VII, seventy-five per cent, and Grade VIII, sixty-eight per cent. The average of these percentages is 62.7. Thus the average pupil in his entire spelling course knows how to spell nearly two-thirds of the words in the spelling list a month or more[1] before he is asked to study them. The Test-study plan requires three reviews of the two-thirds of the words a pupil knows—in the Monday, Wednesday, and Friday tests—and conserves all the rest of the time for mastery of the one-third of the words which he does not know. Theoretically, the Test-study plan is undoubtedly well conceived to make the most of the pupil's time. It should be realized, moreover, that a test may be an effective means of learning. Indeed, in studies of memorizing nonsense syllables and other materials a combination in which a self-test (like recall) predominates over mere rereading, proved to be markedly superior to mere review study.[2]

From the theoretical point of view, the surprising thing—to the writer at least—is the fact that the Pre-test plan does not show to greater advantage than it does in this study and others. While it is neither a perfect nor a fool-proof plan, its disadvantages seem fewer and less serious than those of the Pre-study program. In the writer's opinion most of the limitations of the Pre-test plan previously mentioned are not very serious ones, and most of these may be removed. The most conspicuous deficiency of the plan, according to observations of the writer and of others, is to be found in weaknesses in the pupil's method of study and of management of his own work and the inadequacy of the teacher's supervision of pupil's individual work. The writer's observations were that in the Study-test plan the pupils were held to the use of better techniques of learning, to better distribution of time on different words, to more adequate check-up of results. In the Test-study plan pupils were given less adequate guidance, and often supervision was superficial. As a result, pupils frequently dawdled or more commonly utilized poor methods of study, failed to do their assignment properly, giving undue time to certain words and insufficient time to others, failed to check their work properly, and otherwise relaxed into inferior study. This was notably true of the least experienced and least intelligent pupils, who were precisely the ones among whom the method showed relatively the poorest results.

---

[1] Since the lists used in these tests covered ten weeks of advance assignments, the average would be about five weeks.

[2] Gates, A. I.: Recitation as a Factor in Memorizing. *Archives of Psychology*, 1917.

THE NATURE OF INTELLIGENCE*

J. H WILSON

INTRODUCTORY

In the discussion of the nature of "general intelligence" and of the possibility of testing it, an important suggestion has lately been made by Thorndike. He had put forward a list of eight tests, based apparently on his own theory, and has implied that they will not satisfy the well-known criteria deduced by Spearman and his collaborators for demonstrating the existence of "general intelligence" as a "central factor."

The issue is of much theoretical value and certainly seems important enough to justify putting it to the test. A small research has therefore been planned with this object in view, and the results are described in the following pages.

THE CONDITIONS GOVERNING THE EXPERIMENT

Thorndike's list of tests consists of the following well-known processes: Memory for digits, pitch discrimination, opposites, defining words, completing sentences, arithmetical problems, number series, and completing pictures.

He proposes that tests of this kind be given to 10,000 sixteen-year-olds, the accuracy being such as to secure reliability coefficients of 95. Spearman accepts the tests, but adds that the subjects must be of the same sex, that they should have received reasonably similar education, and that all responses to one and the same test must be uniformly marked by the same person.

As put forward here the proposal is obviously beyond the power of any one investigator. Spearman suggests, however, a somewhat less rigorous procedure which makes it possible to test his prediction First he maintains that such high reliability coefficients are not indispensable. Secondly he suggests that a number of independent workers might each examine about one hundred pupils.

---

* This study has been carried out under the auspices of the Brighton and Hove Higher Education Council. It composed part of a thesis submitted for the Ph D. degree of London University. The author is greatly indebted for valuable assistance to Professor C. Burt and also to the teachers and students who helped in the examination

THE TESTS, SUBJECTS AND ADMINISTRATION

In carrying out the present investigation, use was made of published material wherever possible. Thus memory for digits was tested as described in Whipple's "Manual of Mental and Physical Tests" under "auditory" and "visual" memory respectively. Pitch discrimination was readily measured by Seashore's apparatus and procedure. The verbal tests and the arithmetical ones gave more difficulty. Items were collected from published scales; but the best selection for pupils of sixteen years had to be discovered by special experiment. They were finally chosen after trial on the students of a training college for teachers. A test of completing pictures was to hand in that used in the American Army tests, namely "A Day in the Life of a Schoolboy."

To avoid any influence there may be in the day of examination the tests were prepared in duplicate and given on two days, and to avoid any constant factor entering the results the children changed seats from time to time throughout the examination.

The tests were given to some seventy-odd boys ranging in age from fifteen and one-half to sixteen and one-half, the average being exactly sixteen years and one month. The boys were taken from parallel classes of the same form of a grammar school, the form being the one preparing for the General School Certificate Examination. In consequence of these limitations they formed a highly selected group.

Each test was uniformly marked by the same person. In certain cases this marking was done by Miss D. King of Brighton, in others by the writer, who takes this opportunity of thanking her for such valuable help. Sixty of the pupils had taken both examinations and their results were used

AVERAGE SCORES, STANDARD DEVIATIONS AND RELIABILITIES

These values are given in Table I.

The first point to be considered is the value of each reliability coefficient. The majority of the coefficients lie between .50 and .70. In comparison with .95, the figure suggested by Thorndike, these results are disappointing. Considering, however, how highly selected the group is in age and educational attainments, they compare favorably with those of other workers in allied fields.*

---

* Working with students of university rank, Hazlett obtained values which ranged from .56 down to .20. The writer with younger pupils, using equivalent tests from National A and Terman Group Test of Mental Ability, .37; from Otis and National A, .64, and Terman and Otis, .72

The table also includes the multiples of the present length of each test required to give a reliability equal to that demanded by Thorndike.

Table II gives the intercorrelations of the tests worked out by the usual product-moment formula. These, it will be noticed, are small, for the majority lie between .40 and 0. In large measure this was to be expected on account of the great homogeneity of the examinees. The negative values are not significant.

TABLE I.—AVERAGES, STANDARD DEVIATIONS, AND RELIABILITIES

Test		Average		Standard deviation		Reliability	Multiples to give reliability .95
Name	Number	Form 1	Form 2	Form 1	Form 2		
Memory for digits	1a	7.90	6.52	.952	.901	.49	18
(a) Auditory; (b) visual	1b	11.38	13.50	1.424	3.[illegible]	.65	10
Pitch discrimination	2	27.23	24.22	6.716	6.[illegible]	.78	5
Opposites	3	12.98	11.53	3.[illegible]	4.[illegible]	.57	11
Supplying words	4	16.82	12.45	4.[illegible]	3.[illegible]	.50	13
Defining words	5	11.50	10.26	2.5[illegible]	3.[illegible]	[illegible]	17
Number series	6	12.75	4.03	3.330	4.[illegible]	.72	12
Arithmetic problems	7	10.80	7.60	4.070	3.[illegible]	.73	7
Picture completion	8	79.32	43.27	13.730	15.[illegible]	.55	7

TABLE II.—INTERCORRELATIONS

	Test		1a	1b	2	3	4	5	6	7	8
Memory...	Memory	1a		.545	.211	.454	[illegible]	[illegible]	[illegible]	[illegible]	.366
	Memory	1b	.335		.162	.376	.130	[illegible]	[illegible]	[illegible]	.232
Verbal.....	Pitch	2	.011	−.162		.11	[illegible]	[illegible]	[illegible]	[illegible]	.001
	Opposites	3	.454	.376	.118		[illegible]	.61	[illegible]	.271	.203
	Supplying words	4	.230	.154	−.0[illegible]	.627		.4[illegible]	[illegible]	[illegible]	.154
	Defining words	5	.357	.078	−[illegible]	.616	[illegible]		.271	[illegible]	.127
Arith.....	Number series	6	.206	.003	[illegible]	[illegible]	.31	.2[illegible]		.255	.125
	Arithmetic problems	7	.209	.017	−.076	.171	[illegible]	[illegible]	[illegible]		.154
	Picture completion	8	.304	.23	.064	[illegible]	.1[illegible]	.327	[illegible]	[illegible]	

SPEARMAN'S CRITERIA AND PREDICTIONS

The criterion is as follows: Take any four tests and let the four selected be denoted by the letters *a*, *b*, *p*, and *q*. Then, if the coefficients of correlation are determined solely by a central factor,

$$F = r_{ap}r_{bq} - r_{bp}r_{aq} = 0$$

to the degree that should be expected from the sampling "probable errors" involved. The quantity F is termed the "tetrad difference."

APPLICATION OF THESE CRITERIA

There are three hundred seventy-eight different values of F. Each of these values has been computed, and the distribution of the results is shown in Table III. From this table it is readily seen that many values differ from zero. And, at once there arises the question whether such differences are significant

To answer this question the probable error of the quantity F must be known. This has been found by Spearman and Holzinger *

TABLE III.—TETRAD DIFFERENCES ($N = 3nC_4 = 378$)

RANGE	FREQUENCY
.000-.021[1]	85
.021-.063	148
.063-.105	83
.105-.147	34
.147-.189	21
.189-.231	0
.231-.273	1

[1] The distribution has been made in this way to facilitate the construction of Fig. 1. This curve is then symmetrical.

To compare the magnitude of each "tetrad difference" with its probable error, the expedient has been adopted of dividing the

---

* Where

$$F = r_{12} \cdot r_{34} - r_{23} \cdot r_{14}$$

the probable error of F is

$$.6745\left[\frac{1}{N}\{r^2_{12} + r^2_{34} + r^2_{23} + r^2_{14} - 2(r_{12} \cdot r_{13} \cdot r_{23} + r_{12} \cdot r_{14} \cdot r_{24} + r_{34} \cdot r_{13} \cdot r_{14} + r_{34} \cdot r_{23} \cdot r_{24}) + 4r_{12} \cdot r_{34} \cdot r_{23} \cdot r_{14}\} + \frac{1}{N^2}\{(1 - r^2_{12})^2(1 - r^2_{14})^2(1 - r^2_{23})^2(-r^2_{24})^2\}\right]^{1/2}.$$

The difference between this and the true value is made up of two expressions involving the fourth orders of the standard deviations of the correlation coefficients. These are negligibly small except for small values of the correlation or small numbers of pupils and may be calculated by means of the formula of Filon and Pearson (*Phil Trans. Royal Soc*, London, A CXCI, p 202). Here, however, they are negligible

Even so, the formula remains very cumbersome and it has been customary to use in its place certain approximations This procedure will not be followed here because one of the main difficulties experienced in the theory of two factors has been the use of substitute criteria for the true values. On the contrary, the full formula has been used former by the latter These quotients are summarized in Table IV. In simple sampling a value for the quotient just exceeding 1.00 is as likely to occur as not: a value just exceeding 2.00 is likely to occur about eighteen times in one hundred trials. a value just exceeding 3.00 is likely to occur about one in twenty-three: and for a value just exceeding four, one in one hundred forty-three trials. Until a quotient exceed four there can be no great confidence that it is likely to be "significant." A value of five is usually required in statistical work. A value of three is, however, considered "suggestive."

In Table IV twenty-two of the quotients are greater than three and one is greater than five. These large values constitute a small proportion of the total, and it is reasonable to ask whether they could be expected on the basis of sampling errors.

TABLE IV.—VALUES OF QUOTIENTS $\frac{F}{PE}$

RANGE OF QUOTIENT	FREQUENCY
0–½	80
½–1	70
1–1½	66
1½–2	55
2–2½	50
2½–3	35
3–3½	15
3½–4	6
4–4½	1
4½–5	0
5–5½	0
Greater	1

To decide this question two expedients are available.

In compliance with the first of these devices the values of the quotients have been distributed as is shown in Table V. Column 1 gives the range of the values of the quotients, while Column 2 gives the frequency of their occurrence In Column 4 the frequency which is to be expected, were the distribution normal, is given. Column 5 gives the difference between Columns 2 and 4, and these differences are large.

The probable errors of the differences in Column 5 are readily calculated when the cases entering into the distribution are independent observations. They are not independent in the case of "tetrad differences," for many of these contain the same r's, and the different r's themselves have intercorrelated sampling errors. Values of probable errors calculated on the assumption of the independence of the observations will therefore be too small. They will give, however, a first approximation. Such values are given in the third column of the table, and in the last are to be found the quotients obtained by dividing the value in Column 5 by that in Column 3. Examination of these values makes it highly probable that the distribution of the "tetrad differences" is not normal.

TABLE V.—SHOWING VARIATION FROM NORMALITY OF TABLE VI

Quot $= \frac{F}{PE}$	Frequency	Probable error	Frequency expected	Difference	Quot $= \frac{\text{Diff}}{PE}$
0-1	150	6 5	189	−39	−6 0
0-2	271	5 9	310	−39	−6 6
0-3	356	3.1	362	6	2 0
0-4	376	1 0	375	1	1 0
0-5	377	0 7	378	1	1 4
0-6	378	0	378	0	

Application of the second device supports this conclusion. By means of Table III the histogram shown in Fig. 1 has been constructed In Spearman's "Mental Abilities of Man" (Appendix, p. xi) is given the procedure that is to be followed in order to construct the normal curve to be expected were the differences due to sampling errors alone. This curve is shown in Fig. 1, and it is evident there are variations from what is to be expected by sampling errors.

What, then, is the cause of these perturbations?

Before seeking the causes it is necessary to consider the question at issue To demonstrate the presence of a "central factor" the absence of "tetrad differences" in significant amounts was sought Their presence demonstrates that there is correlation over and beyond that due to the "central factor" On the unifocal theory of Spearman this additional correlation, termed "specific correlation," can only be due to two or more of the mental performances here tested having in common the same "specific" factor.

Further, according to this view, such "group" factors (as they are called) have very narrow incidence

Two methods may be employed in studying this question. First those values of F which are large in comparison with their appropriate probable errors may be considered. Examination of these "tetrad differences" will help to indicate those tests which have "group factors." There are twenty-two values greater than three times the probable errors involved, and eighteen of these, including all greater than four times the probable error, may readily be accounted for by too large values of the coefficients r_{34}, r_{45}, r_{35}, r_{1a1b}, and r_{67}. In particular the coefficient of correlation between the memory tests occurs in four, those between the verbal tests in five, and that between the arithmetical tests in nine, including the greatest value As but one of these values reaches the standard of five times the probable error, it is impossible to do more than say there is a suggestion of the presence of a group factor among the verbal tests, slightly greater suggestion of one between the memory tests, and by far the most evidence in favor of a group factor between the arithmetical tests.

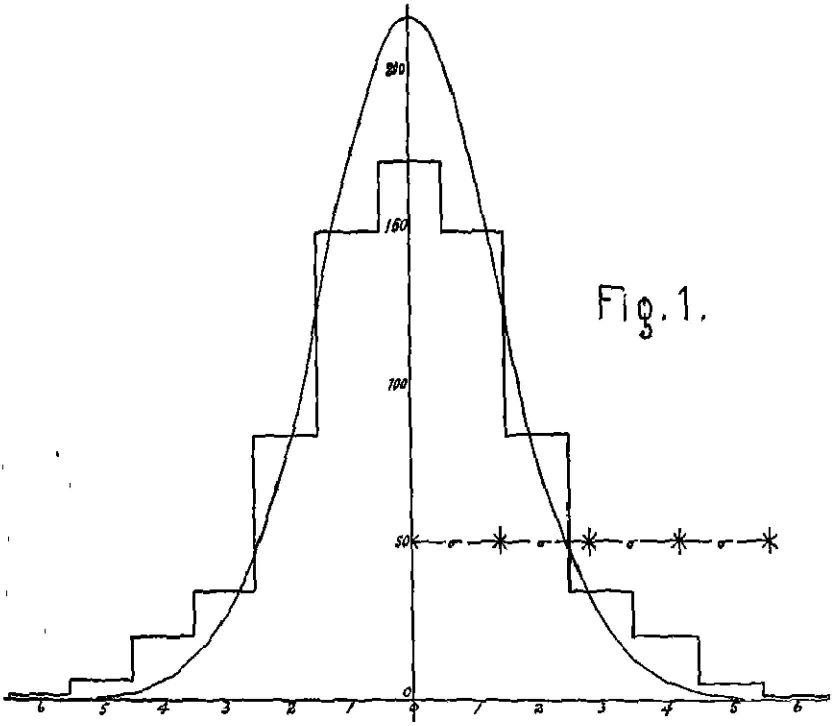


Fig. 1.

Recourse is now had to the second method, in which the amounts of "specific correlation" are computed. Such computation is effected by the use of partial correlation.* To apply the method to a given pair of tests necessitates finding the correlation of each test with the central factor. The most reliable way of obtaining each of these two coefficients is to use all available pairs of tests, but all coefficients of correlation introduced must obey the tetrad equation.† Hence in obtaining the coefficients of correlation of each test with the central factor‡ (see Table VI) no use has been made of the coefficients for the memory tests, nor of those for the arithmetical ones, nor of those for the verbal ones. The values, it will be noticed, are in the main small, the third test (opposites) being the only exception.

TABLE VI.—COEFFICIENTS OF SATURATION

TEST	COEFFICIENT
1*a*	.576
1*b*	.268
2	.040
3	.910
4	.447
5	.675
6	.468
7	.296
8	.387

The coefficients of correlation between the specific factors are summarized in Table VII. In the same table are to be found the quotients obtained by dividing each coefficient by its appropriate probable error. There are few large values for these quotients. Values greater than three occur in the case of the two memory tests, also in the case of defining words and the visual memory tests, again in that of supplying words and opposites tests, in that of arithmetical problems and opposites and finally in the case of the two arithmetical

---

* Write r_{ab} for the correlation between two tests a and b and $r_{Sa,\,Sb}$ for the correlation between the factor specific to a and that specific to b, and there ensues by the formula of Yule for partial correlation, the equation

$$r_{SaSb} = \frac{r_{ab} - r_{ag}r_{bg}}{\sqrt{1 - r^2_{ag}}\sqrt{1 - r^2_{bg}}}$$

where g denotes the factor common to both a and b. For the correlation between the specific factors will be that found when g is held constant. These values have been found for all the possible pairs of tests used.

† Let r_{ag} be the coefficient required, then

$$r^2_{ag} = \frac{r_{ab}r_{ac} + r_{ab}r_{ad} + \cdots + r_{ax}r_{ay} + \cdots}{r_{bc} + r_{bd} + \cdots + r_{xy} + \cdots}$$

a, b, c, etc., being the tests.

‡ These coefficients are technically termed "saturation coefficients."

tests. The only significant value is that obtained for the two arithmetical tests.

A further investigation may now be attempted. Tetrad differences may be selected in the following way:

TABLE VII.—COEFFICIENTS OF CORRELATION BETWEEN SPECIFIC FACTORS

Test		1a	1b	2	3	4	5	6	7	8
Memory	1a		.359 4 7	.230 2 9	−.206 2 1	.003 .0	−.053 .6	−.080 1 0	050 .5	102 2 3
Memory	1b	.359 4.7		− 180 2 1	080 .0	071 0	277 3 1	−.145 1 7	− 067 8	114 1 7
Pitch	2	230 2 9	−.180 2 1		180 2 2	− 003 8	− 153 1 8	230 2 8	− 080 1 0	− 012 1
Opposites	3	− 206 2 1	.080 .0	180 2 2		323 4 1	013 1	− 070 0	258 3 1	.020 3
Supplying words	4	003 0	071 0	−.003 .8	323 4.1		101 2 3	131 1 5	105 1 2	− 059 7
Defining words	5	−.053 .6	.277 3 1	− 153 1 8	.013 1	.101 2.3		− 084 0	.182 2.2	097 1 1
Number series	6	− 080 1.0	− 145 1.7	.230 2 8	−.070 .0	134 1.6	−.084 .0		501 7 7	068 .8
Arithmetical problems	7	.050 .5	−.067 8	−.080 1 0	.258 3.1	.105 1 2	182 2 2	.501 7 7		041 .5
Picture	8	102 2 3	.114 1 7	− 012 .1	020 3	−.059 .7	097 1 1	068 .8	.041 .5	

Take any four tests, including only one at a time from the verbal tests, one only from the arithmetical tests, and one only of the memory tests. The coefficients are determined solely by a "central factor," and the tetrad equation ought to be satisfied to the degree to be expected from the sampling errors.

TABLE VIII.—TETRAD DIFFERENCES PREDICTED AS SATISFYING CRITERIA
($N = 120$)

RANGE	FREQUENCY
000– 015	23[1]
015– 045	39
045– 075	35
075– 105	13
105–.135	8
135– 165	2

[1] The distribution has been made in this way to facilitate the construction of Fig. 2. This curve thus becomes symmetrical.

This may be done by constructing a series of tables similar to Table II, in each of which is placed a selection of four of the tests.

In this way forty such tables may be constructed, and from each table there will ensue three values of the quantity F, making a grand total of one hundred twenty. These values are distributed in Table VIII

TABLE IX · QUOTIENTS = $\frac{F}{PE}$ OF F

QUOTIENT = $\frac{F}{PE}$	FREQUENCY
From 0 to 1	47
Greater than 1 up to 2	49
Greater than 2 up to 3	22
Greater than 3 up to 4	2
Greater than 4 up to ∞	0

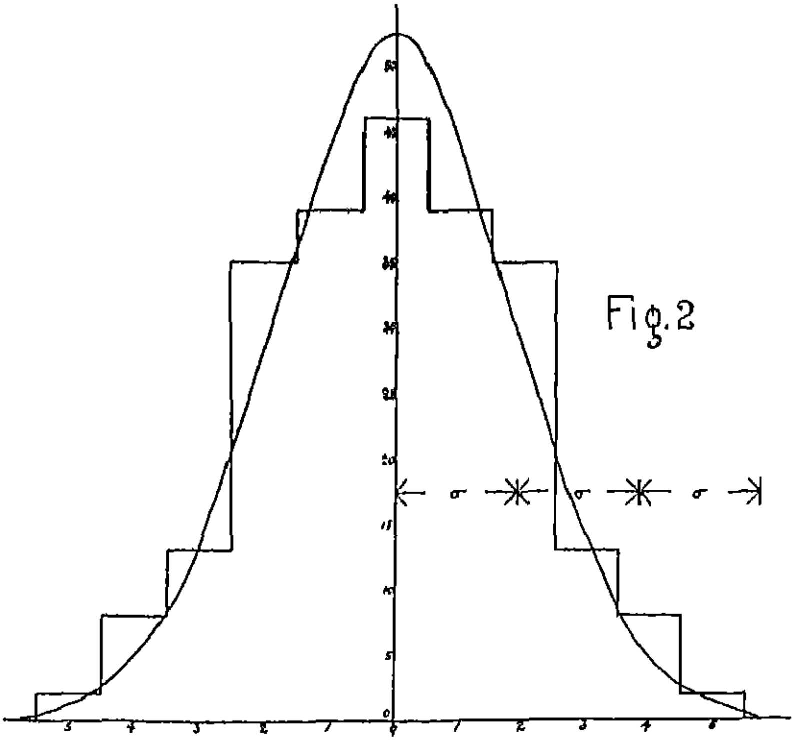


Fig. 2

in a form suitable for the formation of the histogram given in Fig 2. In the same figure is drawn the normal curve to be expected were the differences due to sampling errors. Table IX and Table X have been modeled on lines similar to those employed in Table IV and Table V respectively.

TABLE XIII.—SPECIFIC CORRELATIONS

$N = 50$

	1a	2	3	4	5	6	7	8
1a		+ 104	+.201	– 160	– 165	– 152	108	+ 008
2	+.194		+ 017	– 094	– 151	+ 174	– 030	177
3	+ 201	+ 017		+ 110	–.171	– 062	– 173	125
4	– 160	– 094	+ 140		+ 120	– 019	148	+ 200
5	–.165	–.151	– 171	+ 120		+ 140	+ 032	+ 308
6	– 152	+ 174	–.062	– 019	+ 119		+ 429	- 033
7	– 108	– 030	– 173	– 148	+ 032	+ 429		+ 150
8	+ 008	– 177	– 125	+.200	+ 308	– 033	+ 150	

and opposites. The values are only "suggestive," and among so many values (thirty-six in the first experiment and twenty-eight in the second) can most reasonably be accounted for by sampling errors It is interesting to note, however, that specific correlation has been manifested for boys and not for girls in at least one experiment,[1] though it must be added that in it the tests involved different mental processes

SUMMARY OF THE RESULTS

1. The results obtained with the tests used can readily be explained on the basis of a "central factor" running through all the performances and a number of other factors specific to each performance, provided that the two specific factors belonging respectively to the arithmetical tests overlap.

2. There is, then, conclusive evidence of a "group" factor in the arithmetical tests, and this result agrees with those of Rogers[2] and Collar. Such group factors are of great importance both theoretically and practically They are tacitly assumed in all work on "special abilities" such as "practical ability," "musical ability," and the like. The demonstration of their presence has been made in but few cases The reasons for this lack of experimental evidence are nowhere more succinctly put than by Professor C. Burt.[3] He writes, "Over specific inborn abilities I need not linger. For them effective tests have proved disconcertingly hard to contrive. Simple correlation is here inapplicable General intelligence is always getting in the way. We think we have tested something specific. We find we have only hit upon another test of intelligence. Its ubiquitous influence can only be eliminated by some elaborate technical device, the procedure, for

example, known as multiple correlation; and the complexity of the whole task bewilders where it does not baffle.*

> Nor do these special abilities, although presumably inborn, declare themselves at so young an age as the more general. Specialization during the first twelve years of childhood is the exception rather than the rule . . . the young child contains in fresh and dormant essence the germ of every faculty. Age alone betrays our idiosyncrasies. Adolescence is preeminently the period when many of these localized talents and specialized interests seem for the first time to mature. *Accordingly, efforts at vocational guidance and educational specialization must not be forced at too early a stage.* (Italics mine.) At present, for example, the system of junior county scholarships tends to sweep all our brightest children at the age of ten or eleven into secondary schools of a somewhat academic type. When at a later period examinations are held for trade schools, most of the best instances of special talent are missing: they have already been creamed off and drafted into other directions less suited to their powers.

Among the cases where "group" factors do become of appreciable magnitude the five most important have been in respect of what may be called the logical, the mechanical, the psychological, the arithmetical, and the musical abilities. In each of these a group factor has been discovered of sufficient breadth and degree to possess serious practical consequences—educational, industrial, and vocational.

3. Group factors are evidently absent even where they might otherwise have been expected. There is, for example, no demonstrated group factor in the verbal tests nor one in the memory tests. Particularly evident is the lack of a group factor between the opposites test and that of defining words, for both tests appear to depend largely upon understanding the meaning of words. Yet in neither of the two experiments is there any appreciable specific correlation. The amount of correlation between the specific factors in the memory tests is high but not significant. A similarly high value was obtained by Carey, in which case it was significant.

4. The investigation serves to throw light upon a constituent of the specific factors of an obvious kind and yet not unlikely to possess great importance. It consists in the manner of presentation of the task to the pupils. In the case of the verbal and arithmetical tests the presentation was made by writing. The presentation was oral in the case of the tests of "auditory memory" and pitch discrimination and was both visual and oral in those of "visual memory" and "picture completion." The evidence derived from this investigation would

---

[*] These words were written before the publication of the probable error of the "tetrad difference."

point to there being no specific correlation in the mode of presentation. The choice between oral and written tests, then, would seem to introduce no group factor of appreciable magnitude.

5. The "coefficients of saturation" differ for the different tests and also in the two experiments. For the boys, pitch discrimination contains but very little of the "central factor," while the amount for girls is appreciably higher. A similarly well-marked difference occurs in the opposites test, in the supplying words test, and in the number series test As to the test which gives the best measure of the "central factor," this obviously is the opposites test in the case of the boys and the supplying words test in the case of the girls.

6 To be effective in the evaluation of amount of the "central factor" present in any particular testee's performance, a coefficient of saturation in the neighborhood of .995 is needed. None of the tests used in this investigation even approaches this perfection.

BIBLIOGRAPHY

1. *Vide Psychological Review*, 1922, *Journal of Educational Psychology*, Vol. XV, 1924, pp. 194, 393, *ibid*, Vol. XVI, 1925, p. 423.

2 Macfarlane: *British Journal Psychology*, 1925., *Mon. Supp*, Vol. XIV
Rogers and Collar *Columbia University, Contributions to Education*, No 130, 1923.

3. Burt: "Presidential Address to British Association for Advancement of Sciences, 1923"

ON PARTIAL CORRELATION VS. PARTIAL REGRESSION FOR OBTAINING THE MULTIPLE REGRESSION EQUATIONS

HAROLD D. GRIFFIN

Eureka Springs, Arkansas

THE REGRESSION EQUATION IN EDUCATIONAL PSYCHOLOGY

In the physical sciences, where exact experimental conditions are comparatively easy to maintain, it is possible to predict with considerable confidence the effect of a given set of causes under like conditions In educational psychology, as well as in many other fields, we are unable to isolate factors with such ease. It has therefore been necessary to develop a technique by which one may draw from a mass of data, often conflicting, some conclusion which will represent the most probable relation between variables. This is the correlation technique.

Correlating two variables, however, is but the beginning of serious study. The educational psychologist seeks means of prediction and control For example, no sooner were we measuring intelligence than we were correlating intelligence with school marks, obtaining regression coefficients, and attempting to forecast school success. But as intelligence tests correlate only from 0.40 to 0.65[1] or so with school marks, educators began to use the multiple regression procedures to improve their predictions. Thus, May[2] sought a better prediction of college success by including time spent in study as a third variable. This produced a multiple correlation coefficient ($R_{0.12}$) of 0.824, which was considerably better than the simple correlation (r_{01}) of intelligence and college success, which was 0.60

DIFFICULTIES IN PRESENT PROCEDURES

Multiple regression equations might be employed to a much greater extent were it not for the difficulties, both real and fancied, of the procedures now in vogue There are two general methods of obtaining the multiple regression equation, both seeming to have originated with Yule. One is the partial correlation method, the other is the partial regression method. The former is at best tedious and inadequately checked but, given time and patience, by it one can handle almost any number of variables. The latter solves by means of simultaneous linear equations, but the technique that commonly has been employed, determinants, is very difficult to handle with more than four variables For a three or four variable problem, the partial regression method with any type of solution is much swifter than partial correlation methods. Fortunately, there are methods for solving simultaneous linear equations that are superior to determinants and are capable of handling efficiently any number of variables One such method, the Doolittle, has long been used by the United States Coast and Geodetic Survey and by some civil engineers. The purpose of the present paper is to bring this method to the attention of a larger circle of educational psychologists than at present employ it.

---

[1] A correlation coefficient of 0 40 has but eight per cent forecasting efficiency, and one of 0 65 has but twenty-five per cent See Hull, C. L : The Correlation Coefficient and Its Prognostic Significance. *Journal of Educational Research*, Vol XV, 1927, pp. 327–338; also Wallace and Snedecor; "Correlation and Machine Calculation," Official Publication, Vol. XXIII, No 35, Iowa State College of Agriculture, 1925, p 17.

[2] May, M. A., Predicting Academic Success. *Journal of Educational Psychology*, Vol XIV, 1923, pp. 429–440

HISTORICAL DEVELOPMENT

The possibilities for prediction in the regression equation have been known only to the present generation, and the very concepts of correlation and regression are not much older. An historical survey of the development of these concepts may reveal why procedures in obtaining the multiple regression equation and also efficient methods for solution in intermediary steps still lack standardization.

BRAVAIS AND HIS REPUTED DISCOVERIES

August Bravais, a French geologist and mathematician, was once given credit by Pearson for devising (between the years 1838 and 1846) formulas for solving correlation and intercorrelation between two and three variables.[1] But earlier ideas with regard to Bravais's place in the history of correlation have undergone modification. Pearson himself now holds that Bravais, while working with two and three variables in geodetic work, developed relationships between his product-sums and Gauss's mean-errors (our standard deviations) which, had these relationships been developed, would have led to symbols equivalent to our correlation coefficient. It is doubtful, however, whether Bravais was even thinking of correlations between his observed quantities.[1]

---

[1] Bravais, A. Analyse Mathématique. Sur les Probabilités des Erreurs de Situation d'un Point. *Memoires Académiques de la Royale Scientifique Institute de France, Science, Mathématique et Physique*, Vol. IX, 1846, pp 255–332,

Pearson, K Regression, Heredity, and Panmixia *Phil. Trans. of the Royal Society Ser A*, 187, 1896, pp. 253–318. (See especially pp. 261, 287)

GALTON AND THE CONCEPT "r"

During the middle of the 1870's, Francis Galton was seeking a numerical measure for "reversion." At a lecture delivered at the Royal Institution of Great Britain, February 9, 1877, he presented a symbol for such a measure, and symbolized it "r." Furthermore, he presented it in an equation, $c^2_1 = v^2/(1 - r^2)$, which, of course, may be written $r = \sqrt{1 - v^2/c^2_1}$. In these formulas, v = the variability of a family of sweet peas, and c_1 = the variability of the general population of sweet peas That Galton was clear in his mathematical reasoning may be seen by comparing this formula with Mill's formula for the measure of correlation, $r = \sqrt{1 - S^2_y/\sigma^2_y}$,[2] or with the correlation ratio in its original form, $n_{yx} = \sqrt{1 - \sigma^2_{ay}/\sigma^2_y}$. Galton's lecture was published in *Nature*, Vol. XV, for 1877, where it may be consulted [3]

About ten years later Galton developed an empirical method for determining the correlation between two variables. He made a distribution chart and drew a line across it in such a way as to touch as near the mean of as many rows as possible He then measured the angle of the deviation of this line of best fit from the vertical. The tangent of this angle gave him his index of reversion, or "regression" as he was now terming it. Tangents were used because they swing from +1.00 through 0 to −1.00, thus forming a very convenient measure of varying degrees of deviation from perfect positive to perfect negative. For some time this r was called Galton's function, due to the use of this term by Weldon who applied correlation in his measurement of various sea life.[4]

---

[1] Pearson, K.· Notes on the History of Correlation *Biometrika*, Vol. XIII, 1920, pp 25–45 (See especially pp 28–32)

Darmois, G · "*Statistique Mathématique*," first edition Octave Doin, Paris, 1928 See p 246ff

[2] Mills, F. C., "Statistical Methods Applied to Economics and Business," first edition Henry Holt and Co , 1924, pp 437, 442.

[3] Galton, F : Typical Laws of Heredity. *Nature*, Vol. XV, 1877, pp 492–495, 512–514, 532–533 (See especially pp. 532–533.)

[4] Weldon, W. F. R . Certain Correlated Variations in Crangon Vulgaris *Proceedings of the Royal Society*, Vol LI, 1892, pp 2–21

Weldon, W. F. R. On Certain Correlated Variations in Carcinus Moenas *Proceedings of the Royal Society*, Vol. LIV, 1893, pp 318–320.

EDGEWORTH AND THE TERM "COEFFICIENT OF CORRELATION"

F. Y. Edgeworth (1892) dealt with Galton's function for three variables, and indicated how the method might be applied up to six variables, making the assumption of normal distribution. He used the phrase "coefficient of correlation" instead of reversion, regression, or Galton's function, and the term persisted.[1]

PEARSON AND THE PRODUCT-MOMENT r

By 1896 Pearson (Footnote 1, p. 36) had introduced the product-sum method, enabling the coefficient of correlation to be calculated without the use of an isopleth or thread. Since then we commonly speak of the Pearson product-moment r. Many interesting variations of the Pearson product-moment formula have arisen since its first statement. Symonds has listed some fifty or more variants.[2]

YULE'S TWO METHODS IN MULTIPLE CORRELATION

Partial Regression Method.—G. U. Yule, then an assistant of Pearson, gave a discussion of the product-sum methods in their applications to correlation in two papers published during 1897.[3] In his paper on the theory of correlation Yule obtained partial regression coefficients in a three variable problem by solving an observation equation set up by the method of least squares, using determinants. Merriman, whose "Textbook on Method of Least Squares"[4] (sixth edition) Yule used for setting up his equation, suggests the Gauss direct method of solution (the basis of the Doolittle method) on pages 51–65 of the edition used by Yule, but the hold of the determinant method of solution seems so strong on British mathematicians that neither Yule nor his immediate followers profited by Merriman's suggestions We thus find that our newly revived partial regression method was really the earlier method for obtaining relations between three or more variables. Failure to appreciate the use that could be made of the coefficient of multiple correlation prevented the development of the partial regression method for some years. As late as 1906 Yule could write,[1] "No practical use has, we believe, been made of [the coefficient of multiple correlation] . . . but it appears to have considerable importance, and may be indicative of the closeness of the causal connection between one variable and the joint influence of two other variables of which the first is a function."

---

[1] Edgeworth, F. Y. On Correlated Averages *Philosophical Magazine*, Fifth Series, Vol XXXIV, 1892, pp 190–204

[2] Symonds, P M. Variations of the Product-Moment (Pearson's) Coefficient of Correlation. *Journal of Educational Psychology*, Vol XVII, 1926, pp. 458–469.

[3] Yule, G. U.: On the Significance of Bravais' Formulae for Regression, Etc., in the Case of Skew Correlation. *Proceedings of the Royal Society*, Vol. LX, 1897, pp. 477–489.

Yule, G U.: On the Theory of Correlation. *Journal of the Royal Statistical Society*, Vol. LX, 1897, pp 812–854.

[4] Merriman, M.: "Textbook on Method of Least Squares." New York: John Wiley and Sons. (Many editions, the eighth is of 1911.)

Partial Correlation Method.—It was this pursuit of causal connections that diverted the partial regression method. Yule had been attracted to the possibilities in partial correlation for determining causal relationships. Consequently he developed a method for solving partial r's that leads indirectly toward the coefficient of multiple correlation, but is very laborious. As yet he had neither named nor symbolized the regression coefficient. Yule was a careful investigator and a painstaking reporter; therefore his writings carried the weight of authority, and the methods and terminology that he preferred became the accepted standards for other statistical workers. In 1907 Yule presented the system of notation substantially as used today.[2] It was at this time that b was introduced as the symbol for the regression coefficient, and that a system of subscripts was developed to represent dependent and independent variables. Yule first published his "Introduction to the Theory of Statistics"[3] in 1910–1911. This book has been exceedingly popular and has run into many editions. It has served to codify Yule's methods and techniques, so that statisticians who have been interested primarily in obtaining regression equations, yet only incidentally in partial correlation, have patiently developed their equations by his partial correlation technique.[4]

SYSTEMATIZING PARTIAL CORRELATION METHODS

Truman L. Kelley systematized Yule's partial correlation procedure in 1914 so that one needed to find, for example, but seventy-eight partial r's and one multiple r in a six-variable problem if one were seeking the regression equation[1] Yule's complete method would require two hundred forty partial r's for a six-variable problem. In 1917 Curt Rosenow made a further analysis,[2] reducing the number of partial r's to forty-five in a six-variable problem, but requiring four incidental multiple r's in the process. C. L. Huffaker's schema[3] requires forty-six partials, but only two multiples—one of these merely serves as a check on the other. The schema for three-, four-, and five-variable problems employed by Henry Garrett[4] resembles Huffaker's. J. E. Bathurst has extended Huffaker's schema to include seven and eight variables[5] Yule's complete method requires three hundred fifteen partials for a seven-variable problem, and five hundred eighty-eight for an eight-variable. Bathurst requires eighty-three and one hundred twenty-nine respectively. In 1916 Kelley contributed a cleverly worked out set of tables to assist in solving partial correlations.[6] The first edition of this was soon exhausted, however, and, for various reasons, it has never been reissued.

---

[1] Hooker, R. H. and Yule, G. U. Note on Estimating the Relative Influence of Two Variables upon a Third. *Journal of the Royal Statistical Society*, Vol. LXIX, 1906, pp. 197–200.

[2] Yule, G. U.: On the Theory of Correlation for Any Number of Variables Treated by a New System of Notation. *Proceedings of the Royal Society* (*Series A*) Vol. LXXIX, 1907, pp. 182–193.

[3] Yule, G. U.: "Introduction to the Theory of Statistics," seventh edition, revised. Charles Griffin and Co., London, 1924.

[4] Yule: *Ibid.*, pp. 226–248.

RETURN TO PARTIAL REGRESSION METHODS

In the herculean task of gathering the data and publishing the results of the Army psychological examinations a more direct method was required for finding the regression equations Brown and May, at the suggestion of Karl Pearson and Raymond Pearl, used simultaneous linear equations with solution by determinants,[7] as Yule had done nearly a quarter century earlier. Kelley then turned his attention to methods for shortening the determinant solution of the simultaneous linear equations. In 1921 he published a nomogram,[1] for facilitating computation of the partial regression coefficients. That same year Clark L. Hull published a description of a nomographic method for solving partial correlations,[2] but his alinement board was suggested by the partial correlation technique and Kelley's "Tables" of 1916 rather than by the partial regression method. In 1923 E. R. Wood prepared nomograms for solving the formulas for partial correlation coefficients and the formulas for partial regression coefficients. Wood's charts constitute by far the best graphic solutions for these two formulas yet proposed, but unfortunately they have not yet been published commercially.[3] Symonds, following Kelley, made a job-analysis of determinant solutions in three- and four-variable problems, and published charts based on this method.[4] But solution by determinants proves troublesome beyond four variables.[5]

---

[1] Kelley, T. L: "Educational Guidance," (first edition), Teachers College, Columbia University Contributions to Education, No. 71, 1914

[2] Rosenow, C: "The Analysis of Mental Functions." Psychological Monographs, Vol XXIV, No. 5, 1917

[3] Huffaker, C L. A Contribution to the Technique of Partial Correlation. *Journal of Applied Psychology*, Vol VII, 1923, pp. 135-142

[4] Garrett, H. E. "Statistics in Psychology and Education," first edition. Longmans, Green and Co., 1926, pp. 223-231, 240-251

[5] Bathurst, J. E: A Partial Correlation Schema. *Journal of Applied Psychology*, Vol. XI, 1927, pp. 155-164.

[6] Kelley, T. L.: Tables: "To Facilitate the Calculation of Partial Coefficients of Correlation and Regression Equations." *Bulletin* No. 27, University of Texas, 1916

[7] Yerkes, R. M.: "Psychological Examining in the United States Army." Memoirs of the National Academy of Sciences, Vol XV, Government Printing Office, 1921

---

[1] Kelley, T. L "Chart to Facilitate the Calculation of Partial Coefficients of Correlation and Regression Equations," first edition. School of Education, Special Monograph, No 1, Stanford University Publications, 1921

Kelley, T L : "Alignment Chart of Correlation Functions." Stanford University Publications, 1921

See also, Kelley, T. L : "Statistical Method," first ed. The Macmillan Co., 1923, pp 291–295 and inside back cover.

[2] Hull, C. L : A Device for Determining Coefficients of Partial Correlation. *Psychological Review*, Vol. XXVIII, 1921, pp. 377–383.

[3] Wood, E. R.: "A Chart for Obtaining Partial Correlations and Regression Equations of Three or More Variables." To be issued by the University of Chicago Press.

[4] Symonds, P. M : Job-analysis Sheet for Computing Partial and Multiple Coefficients of Correlation and Regression Coefficients. *Teachers College Record*, Vol. XXVII, 1925, pp. 52–60.

Symonds, P. M.: "Partial and Multiple Correlation Chart. Three Variables." Teachers College, Columbia University, Bureau of Publications.

Symonds, P. M.: "Partial and Multiple Correlation Chart Four Variables" Not now listed for sale.

[5] For methods of solution by determinants see, Whittaker, E. M and Robinson, G.: "The Calculus of Observations," second edition. Blackie and Son, London, 1926, pp. 71–77, 231–234. The first reference is to Chio's method of solution

Deming, H G : A Systematic Method for the Solution of Simultaneous Linear Equations. *American Mathematical Monthly*, Vol XXXV, 1928, pp 360–363. An orderly application of Chio's method.

Hanus, P. H.: "Elementary Treatise on The Theory of Determinants," first edition. Ginn and Co., 1886. Largely follows the methods and practice of Muir.

Salmon, G.: "Lessons Introductory to the Modern Higher Algebra," fourth edition. Hodges, Figgis and Co., Dublin, 1885. Fourth edition reprint, G E. Stechert, New York, 1924. A thorough introduction to determinants following the methods and practices of A. Cayley and J. J. Sylvester.

ITERATION METHODS OF SOLUTION

There are, however, other ways by which simultaneous linear equations may be solved. Carr[1] lists some six or seven direct algebraic methods. There are also certain indirect methods of reaching the results by iteration, or successive approximations.[2] Recently, Kelley and Salisbury have advanced another such method[3] which they claim will reduce the labor of computation in a sixteen-variable problem at least ninety-five per cent over determinant methods. Tolley and Ezekiel have shown, however, that the Kelley-Salisbury iteration method in solving a six-variable problem for partial regression and multiple correlation coefficients is far inferior to the Doolittle method.[4] Kelley grants the superiority of the Doolittle method "for a small number of variables—perhaps up to 10," and recommends its use; but he insists that the recent improvements on his iteration method make it the swifter instrument where a great number of variables are involved.[5]

THE DOOLITTLE METHOD

The Doolittle method is a variation of Gauss's direct method of solving simultaneous linear equations by substitution. Gauss developed several methods of solution—some direct, others indirect.[1] M. H. Doolittle of the United States Coast and Geodetic Survey made various improvements on Gauss's method of direct substitution, and in 1878 published an account of his method.[2] Doolittle's method replaced Schott's version of Gauss's approximation method in the work of the coast and geodetic survey,[3] and found favor among engineers.[4] In 1923, Howard R. Tolley and Mordecai M. B. Ezekiel of the United States Bureau of Agricultural Economics introduced this method for the solution of the partial regression coefficients (Kelley's β's) into statistical practice,[5] in which the writers adopted the novel method of using the mean product sum, $p_{01} = \Sigma_{01}/N$, etc., in the simultaneous linear equations, instead of the zero-order coefficients as is the general practice. Mills adopted the method as there presented for his text on economic and business statistics published the following year.[6] Hull also uses the entire method as outlined by Tolley and Ezekiel in his 1928 book, "Aptitude Testing."[7] In 1925 Wallace and Snedecor used the method, but with zero-order coefficients instead of mean product sums, in their little manual for agricultural research workers.[1] Garrett, 1928, saw that zero-order coefficients could be substituted for the mean product sums, but he evidently was unfamiliar with Wallace and Snedecor's study, and he also seems to have failed to investigate Tolley and Ezekiel's reference to the Doolittle method of solution.[2] The writer would recommend Wallace and Snedecor's study to educational psychologists as the clearest and most adequate presentation of the Doolittle method now available.

---

[1] Carr, G. S · "A Synopsis of Elementary Results in Pure Mathematics." Francis Hodgson, London, 1886, pp. 42ff.

[2] Kelley, T. L.: "Statistical Method." Pp. 302–310.

Whittaker and Robison: *Op. cit.*, pp. 255ff.

Edgeworth, F. Y.: A New Method of Reducing Observations Relating to Several Quantities. *Philosophical Magazine* Ser 5, Vol. XXIV, pp. 222–223, 466–479; Vol. XXV, 1888, pp. 184–191.

Ford, L R.: The Solution of Equations by the Method of Successive Approximations. *American Mathematical Monthly*, Vol. XXXII, 1925, pp 272–287.

[3] Kelley, T. L and Salisbury, F. S : Iteration Method for Determining Multiple Correlation Constants. *American Statistical Association*, Vol XXI, 1926, pp. 282–292.

Salisbury, F S.: "A Simplified Method of Computing Multiple Correlation Constants. *Journal of Educational Psychology*, Vol. XX, 1929, pp. 44–52. An improvement on the iteration method explained in the preceding article.

[4] Tolley, H. R and Ezekiel, M. M. B · The Doolittle Method for Solving Multiple Correlation Equations *vs.* the Kelley-Salisbury Iteration Method. *American Statistical Association*, Vol. XXII, 1927, pp. 497-500.

[5] Kelley, T. L. and McNemar, Q.: Doolittle *vs* the Kelley-Salisbury Iteration Method for Computing Multiple Regression Coefficients *American Statistical Association*, Vol. XXIV, 1929, pp. 164–169.

---

[1] For some of Gauss's methods see, Merriman: *Op. cit*, eighth edition Pp. 181–187.

Whittaker and Robinson: *Op. cit.*, pp. 234–236, 257–258.

Encke, J. F.: Ueber die Methode der Kleinsten Quadrate. *Astronomische Jahrbuch*, Berlin, 1835, pp. 267–272; 1836, p. 265 Gauss's direct process for solution by elimination.

Jacobi, K.: "Astronomische Nachrichten," Altona, No. 523, 1845, p. 297. Gauss's method of successive approximations.

Schott, C. A.: Solution of Normal Equations by Indirect Elimination. *Coast Survey Report*, 1855, pp. 255–264. Gauss's indirect process of elimination by successive trials and approximations revised and systematized for use in the coast survey. This method was largely used there for twenty-five years until replaced by the Doolittle method

[2] Doolittle, M. H.: Method Employed in the Solution of Normal Equations and the Adjustment of a Triangulation. *Coast and Geodetic Survey Report*, 1879, pp. 115–120

[3] See, Wright, T W. and Hayford, J. F · "Adjustment of Observations by Methods of Least Squares," second edition D Van Nostrand Co , New York, 1906, preface

Also, Adams, O S · Application of the Theory of Least Squares Special Publication No 28. *Coast and Geodetic Survey*, 1915.

[4] Leland, O. M.: "Practical Least Squares," first edition McGraw-Hill Book Co., New York, 1921.

[5] Tolley, H. R. and Ezekiel, M. M. B.: A Method of Handling Multiple Correlation Problems. *American Statistical Association*, Vol. XVIII, 1923, pp 993–1003.

[6] Mills: *Op cit*, pp. 491ff., 576–581.

[7] Hull, C L., "Aptitude Testing," first edition. World Book Co , 1928

CONCLUSION

Thus we have seen that the methods for obtaining the regression equations and the coefficient of multiple correlation were hampered for many years by inadequate and cumbersome methods of solution. Now that a synthesis of the most economical statistical method for obtaining the regression equations, the partial regression method, has been effected with the most economical engineering method for solving simultaneous linear equations, the Doolitte method, we may expect that the multiple correlation and prediction technique will be employed to a much greater extent in educational psychology than in the past.

---

[1] Wallace, H. A. and Snedecor, G. W., "Correlation and Machine Calculation" Official publication, Vol. XXIII, No. 35, Iowa State College of Agriculture, 1925.

[2] Garrett, H. E: A Modification of Tolley and Ezekiel's Method of Handling Multiple Correlation Problems. *Journal of Educational Psychology*, Vol. XIX, 1928, pp. 45–49.

THE SHRINKAGE OF THE COEFFICIENT OF MULTIPLE CORRELATION[1]

SELMER C. LARSON

Carleton College, Northfield, Minnesota

INTRODUCTION

It has been recognized by theoretical statisticians for some time *that when the coefficient of multiple correlation (R) is derived for a* given set of data, its value is likely to be deceptively large. If the computations have been correct, the value will hold rigidly for the set of data from which the regression equation was derived. If, however, the equation should be applied to a second set of data, even though strictly comparable, it has been supposed that the yield in this latter case would, except for errors due to sampling, be less than in the first. Moreover, it has been supposed that the more variables contained in the regression equation, the greater this shrinkage will be. This is particularly significant because ordinarily the practical employment of a regression equation involves its use with data other than those from which it was derived. If this shrinkage should turn out to be very large, the building of multiple regression equations might well be abandoned. The matter is therefore one of considerable importance, both theoretically and practically. Several attempts have been made by statisticians to derive a formula which should indicate the amount of this shrinkage. The most promising one of these will be considered *in the present paper. So far as the writer has been able to discover,* no one has attempted to determine experimentally the actual amount of shrinkage. The present report describes such an attempt in the field of psychological testing.

A study of the shrinkage is made by using a regression equation derived from one group of subjects to predict the criterion scores of a second group. The correlation yield by this procedure is subtracted from the yield obtained by predicting the criterion scores of the second group by means of a regression equation derived from themselves. This shrinkage is studied with the number of the independent variables in the regression equation ranging from one to ten in number and for a variety of different criteria. A comparison is then made between such empirical findings and the results obtained by applying a recently proposed formula for determining the same type of shrinkage.[1]

---

[1] From the Psychological Laboratory, University of Wisconsin. The writer is greatly indebted to Professor M. V. O'Shea for permission to use data selected from the results obtained by the Mississippi Survey

SOURCE AND SELECTION OF DATA

Something like 30,000 pupils were given mental and achievement tests in a survey of the school system of the State of Mississippi. For the present study, the test scores of eight hundred high school pupils from this number were used. The scores of pupils from the large and medium-sized high schools were chosen in the belief that the level of instruction would be more nearly uniform.[2] The entire population of eight hundred consisted of four groups—two hundred boys in each of two groups and two hundred girls in each of the two remaining groups. The subjects to make up these groups were chosen from those tested by the survey in such a way that each of the four contained exactly the same number of individuals drawn from any particular class of each school sampled. Otherwise the placing of the subjects in the several groups was entirely at random. By making up the personnel of the groups in this manner it was felt that they would be as exactly comparable in regard to general level and range of natural endowment, culture, and educational opportunities as possible.

Each pupil had eighteen scores entered after his name. The designations are as follows.

X_1 English
X_2 Mathematics
X_3 Science
X_4 History
X_5 Chronological age
X_6 Intelligence quotient
X_7 Information (Terman)
X_8 Best answer (Terman)
X_9 Word meaning (Terman)
X_{10} Logical selections (Terman)
X_{11} Arithmetic (Terman)
X_{12} Sentence meaning (Terman)
X_{13} Analogies (Terman)
X_{14} Mixed sentences (Terman)
X_{15} Classifications (Terman)
X_{16} Number series (Terman)
X_{17} Total Terman
X_{18} Total Iowa

The first four X's together with X_{18} are scores made on the Iowa High School Content Examination. X_7 to X_{17} are scores made on the Terman Group Test of Mental Ability. Another column—the sum of each row—was added for checking purposes.

---

[1] Ezekiel, M. J. B. An unpublished paper read before the Mathematical Society at its annual meeting in Chicago in December, 1928.

[2] O'Shea's study showed that for the state as a whole scholastic achievement in the small high schools was decidedly lower than in the larger ones.

EMPIRICAL DETERMINATION OF SHRINKAGE

PART I

Ten distinct regression equations with English as the criterion were derived from the data for the first group of boys. Ten parallel regression equations were derived from the data for the second group of boys. In the case of the first group, the first equation had all ten independent variables (tests) The second equation had the best nine test variables, *i.e.*, those having the highest criterion correlations. The third had the best eight test variables, and so on down to the tenth equation which was based on the single test having the highest criterion correlation. The same procedure was followed with the second group of boys, except that in this case the same test variables were used in the corresponding equations as were used with the first group. Owing to a natural variability in the size of the zero order correlation coefficients from sample to sample, the tests successively excluded from the progressively smaller equations with the second group of boys were not in all cases the next in order of weakness in the criterion r's

Space is lacking for the presentation either of the means and standard deviations needed for the derivation of the regression equa-

TABLE I —ZERO ORDER COEFFICIENTS OF CORRELATION FOR BOTH SETS OF BOYS

The Bold Face Figures at the Upper Right Are the Coefficients of Correlation for Boys, Set No II; and the Light Face Figures at the Lower Left Are the Coefficients for Boys, Set No I At the Top and Left of Table Are Indicated the Variables Whose Notations Are Outlined Earlier in the Text This Table Shows How English (X_1) Correlates with Each of the Items in the Terman Test and Also the Intercorrelations between the Various Items

	x_1	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}
x_1	1 000	**.680**	**.533**	**.740**	**531**	**.367**	**502**	**.456**	**.418**	**.489**	**204**	**.718**
x_7	.600	1 000	**.714**	**697**	**631**	**425**	**565**	**.454**	**.451**	**.516**	**.204**	**812**
x_8	490	617	1 000	**602**	**.617**	**.447**	**.522**	**.433**	**.439**	**442**	**.349**	**.786**
x_9	704	621	460	1 000	**587**	**348**	**.587**	**.446**	**.527**	**531**	**.247**	**837**
x_{10}	521	612	408	580	1 000	**384**	**466**	**.460**	**.391**	**434**	**286**	**737**
x_{11}	306	310	272	302	382	1 000	**.280**	**.406**	**.210**	**366**	**416**	**.612**
x_{12}	506	177	411	507	407	382	1 000	**.297**	**.392**	**.356**	**204**	**.688**
x_{13}	503	515	402	107	151	438	121	1 000	**345**	**362**	**450**	**653**
x_{14}	400	625	108	502	427	327	500	380	1 000	**.433**	**254**	**.634**
x_{15}	171	510	400	505	470	335	300	190	350	1 000	**288**	**.639**
x_{16}	255	310	331	350	121	553	312	531	280	418	1 000	**.545**
x_{17}	714	705	670	812	730	613	700	728	604	650	.650	1 000

TABLE II.—SHOWING THE ACTUAL SHRINKAGE IN R FROM THE THEORETICAL VALUE OF AN EQUATION DERIVED FROM A GROUP OF SUBJECTS WHEN AN EQUATION DERIVED FROM A COMPARABLE GROUP IS APPLIED TO THEM. THE CRITERION THROUGHOUT IS HIGH SCHOOL ACHIEVEMENT IN ENGLISH

		Number of test variables used in prediction									
		1	2	3	4	5	6	7	8	9	10
Boys' Group No. I											
Correlation yield (R) from equations derived from their own scores	A	7042	7794	7834	7872	7880	7907	7929	7941	7944	7945
Correlation yield (R) from equations derived from scores of Group II	B	7042	7773	7798	7836	7820	7863	7827	7866	7847	7832
Shrinkage	C	0000	0021	0036	0036	0060	0044	0102	0075	0097	0113
Boys' Group No. II											
Correlation yield (R) from equations derived from their own scores	D	7402	7759	7813	7826	7847	7858	7859	7863	7868	7869
Correlation yield (R) from equations derived from scores of Group I	E	7402	7728	7794	7803	7806	7725	7786	7821	7794	7786
Shrinkage	F	0000	0031	0019	0023	0041	0133	0073	0042	0074	0083
Mean shrinkage of both groups $\frac{(C+F)}{2}$	G	0000	0026	0027	0029	0050	0088	0087	0058	0085	0098

tions or for the regression equations themselves.[1] In order that the interested reader may study the relationship between the original correlations and the several multiple correlation yields, there has been placed in Table I the entire series of zero order correlation coefficients for both groups of boys. The coefficients of the respective groups are distinguished by means of contrasting type faces. All of the coefficients are positive.

The multiple correlation coefficients derived from the results shown in Table I are given in Table II. The R's corresponding to the several multiple regression equations derived from the boys of Group I are shown in row A and those for Group II in row D. These values are the correlation coefficients which would have been obtained if in each case the test scores from which the regression equation was derived had been substituted appropriately in the equation itself and the resulting criterion estimates had been correlated with the true criterion in the ordinary way. Actually, these values were obtained by means of the usual formula which is decidedly simpler. The coefficients in both series are distinctly high, as aptitude correlation yields run. It is noteworthy, however, that in both series alike, after three tests have been included in the equation, the addition of all the remaining seven tests suffices to raise the correlation yield a total of barely a single point in the second decimal place.

The next step in the process was to substitute the actual test scores of the boys of Group I in the equations derived from Group II and to correlate the resulting criterion estimates with the true criterion scores of Group I. The resulting series of coefficients is given in Table II, row B. The procedure was then reversed. The true criterion of Group II was correlated with the criterion estimates obtained by substituting their relevant test scores in the equations derived from the scores of Group I. The resulting coefficients are given in row E. We now have the values from which shrinkages may be determined.

According to the *a priori* expectation as indicated above, the values in row B should show a perceptible shrinkage when compared with the values in row A and similarly with row E when compared with row D. A brief comparison of the R values in the two pairs of rows shows that, except for the equation containing but a single test variable, this expectation is realized. The amount of the shrinkage is shown in row *C* for the several pairs of values of Group I, and for Group II in row *F*. The mean values for rows *C* and *F* are given in row *G*. A glance at these latter values shows at once that, despite a certain amount of variability presumably due to sampling errors, there is a clear tendency for the shrinkage to increase with the increase in the number of independent (test) variables in the regression equation. This again is in harmony with what has been believed by statistical theorists. A graphic representation of the mean shrinkage values shown in row *G* is presented as the solid line in Fig. 1

---

[1] These are given in detail for the entire study in the author's dissertation filed in the library of the University of Wisconsin. It is entitled "Studies in Aptitude Forecasting with the Multiple Regression Equation."

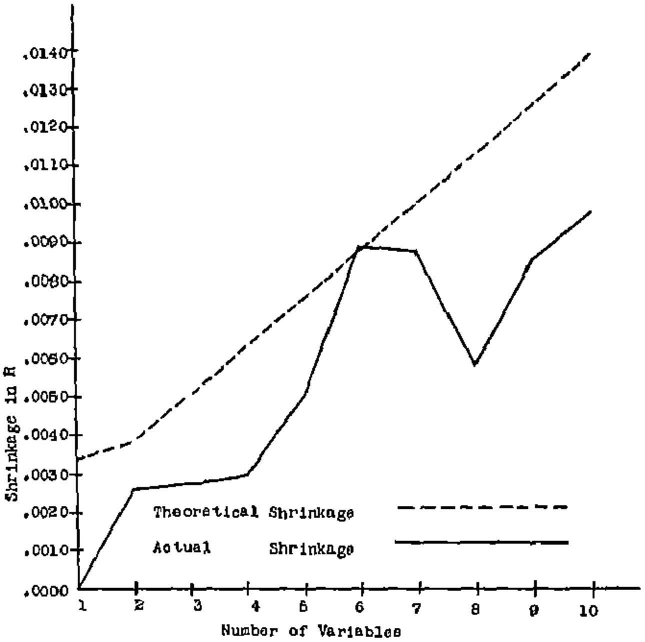


Fig. 1.—Shrinkage as obtained by the use of the formula and also as obtained experimentally.

We have seen that the increase in the size of *R*, even when regression equations are applied to the same data from which they are derived, is extremely slight and grows less and less as the number of test variables increases. We have also seen that under the same condition the amount of shrinkage in yield from such equations when in actual use grows greater and greater. The question naturally arises whether there may not be a point beyond which the increase in *R* resulting from the addition of a new test variable may not be more than offset by the increase in shrinkage, so that the true functional value of such a test battery or other estimating aggregate may not be actually less than if the test or other independent variable had not been added. An examination of rows *B* and *E* shows that in the case of both groups alike such a critical point is reached at the eighth test added. *In both cases the batteries would have given an absolutely higher forecasting yield with two of the tests left out.* The moral of this is that under some circumstances the inclusion of certain tests in an aptitude battery after it has reached some size may not only entail the waste of energy and materials of administering the test, but may actually reduce the yield, and this even when the best possible method of weighting the tests is employed.

PART II

To secure further empirical evidence as to the amount of shrinkage from the ordinary multiple correlation coefficient, further computations were undertaken in all of which the number of test variables was kept constant. The number chosen was the maximum for this study—ten The multiple regression equations on mathematics, science,

TABLE III.—SHOWING THE SHRINKAGE OF *R*'S WHEN THE SAME (10) TEST VARIABLES ARE USED BUT DIFFERENT ACADEMIC SUBJECTS ARE EMPLOYED AS CRITERIA WITH DIFFERENT CORRELATION YIELDS

		English	Mathematics	Science	History	Total Iowa
Boys' Group II.						
Correlation yield (*R*) from equations derived from the subjects' own scores	*A*	7869	6431	5689	7719	8200
Correlation yield (*R*) from equations derived from the scores of Group I	*B*	7786	6148	5230	7505	8098
Shrinkage	*C*	0083	0283	0459	0214	0102
Girls' Group II						
Correlation yield (*R*) from equations derived from the subjects' own scores	*D*	7989	6403	4548	7115	7875
Correlation yield (*R*) from equations derived from the scores of Group I	*E*	7665	6226	4219	6786	7755
Shrinkage	*F*	0324	0177	0329	0329	0120

history, etc., for the boys of Group I were used to estimate the corresponding true criterion scores of Group II. The same procedure was followed for the girls.

The results are given in Table III which is constructed in a manner comparable to Table II above. With the number of test variables constant, this table enables us to observe the influence upon the amount of shrinkage of the strength of the natural tendency to correlation in the test data involved. If we divide the ten shrinkages found in this series into two groups on the basis of the size of the original *R*'s, we find that the average shrinkage for the five largest *R*'s is .0169 whereas that for the five smallest *R*'s is .0315. The tendency for the weaker sets of data to yield the larger shrinkages is evident. For the two lowest values (Science) this amounts to .0394, a very appreciable amount.

THE SMITH SHRINKAGE-DEDUCTION FORMULA

A promising correction formula has been developed to apply to the coefficient of multiple correlation. A paper containing the formula was read by M. J. B. Ezekiel at the December 1928 meeting of the American Mathematical Society held at Chicago. This formula is

$$\bar{R}^2 = 1 - \frac{1 - R^2}{1 - \frac{m}{n}} \quad \text{or} \quad \bar{R}^2 = \frac{nR^2 - m}{n - m}$$

where $\bar{R}$ = the estimated correlation obtaining in the universe
R = the observed correlation
m = the number of variables, dependent and independent
n = the number of observations (statistical population)

Ezekiel gives the credit for developing this formula to B. B. Smith.

At the completion of the computations described in the preceding section it was a relatively simple task to substitute in the above formula and determine for the various observed *R*'s, the corresponding estimates of the "correlation obtaining in the universe." These values are given in Table IV for the *R*'s obtained in Part I. Table V shows the corresponding values of the *R*'s obtained in Part II. Subtraction from the original *R*'s shown in Tables II and III respectively yields the amounts of shrinkage which would be anticipated by the formula in each case.

In Table IV row *E* shows the mean amount of shrinkage for each pair of observed *R* values for the several numbers of test variables.

TABLE IV —SHOWING THE SHRINKAGE OF R'S AS INDICATED BY THE SMITH FORMULA, WHERE THE VARIABLES RANGE IN NUMBER FROM ONE TO TEN

		Number of test variables used in prediction									
		1	2	3	4	5	6	7	8	9	10
Boys' Group I											
Correlation obtaining in the universe ($\overline{R}$) as estimated by the Smith formula.	A	7006	7756	7784	7810	7805	7821	7831	7831	7821	7809
Shrinkage obtained by subtracting the values in row A above from those in row A of Table II	B	0036	0038	0050	0062	0075	0086	0098	0110	0123	0136
Boys' Group II											
Correlation obtaining in the universe ($\overline{R}$) as estimated by the Smith formula	C	7371	7720	7762	7762	7771	7769	7757	7748	7740	7727
Shrinkage obtained by subtracting the values in row C above from those in row D of Table II	D	0031	0039	0051	0064	0076	0089	0102	0115	0128	0142
Mean shrinkage of both groups $\frac{(B+D)}{2}$	E	.0034	0039	0051	0063	0076	0088	0100	0113	0126	0139

It is plotted in Fig. 1 for the purpose of easy comparison with the empirically determined shrinkages. It is clear that the formula definitely parallels the empirical findings It conforms to the observed tendency for the amount of shrinkage to increase with the number of variables involved in the equation There is a well-marked tendency, however, for the Smith formula to indicate materially larger shrinkages than the empirical results show

Passing to Table V, we may make the comparison with the empirical results by treating the shrinkages the same as those in Table III

TABLE V—SHOWING THE SHRINKAGE OF R'S AS INDICATED BY THE SMITH FORMULA, WHERE THE NUMBER OF VARIABLES IS CONSTANT BUT THE SIZE OF THE R'S VARY RATHER WIDELY

		English	Mathematics	Science	History	Total Iowa
Boys' Group II.						
Correlation obtaining in the universe ($\overline{R}$) as estimated by the Smith formula . .	*C*	7727	6156	5330	7563	8094
Shrinkage obtained by subtracting row *A* above from row *A* in Table III . .	*B*	0142	0275	0359	0156	0106
Girls' Group II.						
Correlation obtaining in the universe ($\overline{R}$) as estimated by the Smith formula . .	*C*	7843	6132	4005	6916	7730
Shrinkage obtained by subtracing row *C* above from row *D* in Table III	*D*	0146	0271	0543	0199	0145

Computation shows that the mean shrinkage of the five largest R's is .0139 whereas that for the five smallest is 0329 Here again we observe, as in Table IV, that the shrinkages yielded by the formula are materially larger than those found empirically. An analysis of the formula reveals that the shrinkage increases as the size of the obtained R decreases. The formula will break down, however, when $m/n > R^2$ as the values will then become imaginary In this situation, with m equal to 11 and n equal to 200, the formula will give imaginary values when the absolute values of R are less than 235.

SUMMARY AND CONCLUSIONS

1. The present investigation has shown that the theoretically expected shrinkage of *R* as derived by the multiple correlation formula is a fact.

2. The shrinkage is found to increase as the number of test variables increases.

3. The shrinkage is also found to increase as the size of *R* decreases.

4. The Smith shrinkage-deduction formula parallels all of the above empirical findings, but quite consistently gives values which are in excess of those obtained under the present experimental conditions.

5. The empirically observed shrinkage increases at such a rate with the increase in the number of test variables that one of the most widely known scholastic aptitude tests actually shows a lower correlation yield with a criterion when ten test units are used than when only eight are employed. This suggests that test batteries may have very definite limitations as to size

THE INFLUENCE OF BLOOD RELATIONSHIP AND COMMON ENVIRONMENT ON MEASURED INTELLIGENCE

VERNER MARTIN SIMS

University of Alabama

One of the favorite modes of attack used in the investigation of the relative influence of heredity and environment on intelligence as measured by the tests of today has been through a consideration of the resemblance of siblings (brothers and sisters) in the trait or traits measured by these tests. These studies are numerous[1] and have invariably shown a decided correlation between the intelligence of siblings. Interestingly enough, this correlation coefficient is about the same as the correlation found between the physical characteristics (eye color, cephalic index, stature, etc.) of siblings—approximately .50. However, this similarity between the physical and intellectual resemblance of siblings does not necessarily mean that the relationship is determined by the same causative factors. The physical characteristics are seemingly uninfluenced by environment, but the siblings are certainly subjected to a common environment, and, until evidence is produced that this is a negligible factor in the case of intelligence, the correlation has no significance. In recent years this fallacy has been pointed out time and again. It but remains to attempt the determination of the relative contributions of common environment and blood relationship to the correlation which is found.

One of the most satisfactory approaches made to this problem that has come to the writer's attention is contained in the Chicago study reported in the 1928 Year-book of the National Society for the Study of Education.[2] In this study the correlation between the intelligence of siblings reared in the same home was compared with that of siblings who were separated before either child was six years of age. Instead of the usual correlation of .50 a correlation of .32 ± .05

---

[1] For careful summaries of most important of these studies see Burks, Barbara. A Summary of Literature on the Determiners of the Intelligence Quotient and the Educational Quotient. *Twenty-seventh Yearbook of the National Society for the Study of Education*, 1928, Part II, Chap. XVI, pp. 252–261.

[2] Freeman, Holzinger, and Mitchell: The Influence of Environment upon Intelligence, School Achievement and Conduct of Foster Children. *Twenty-seventh Yearbook of the National Society for the Study of Education*, 1928, Part I, pp. 128–135.

(by age pairing) or .25 ± 06 (by double entry pairing) was found between the intelligence of one hundred twenty-five pairs of siblings who had been separated for an average of seven years and four months and had an average age of five years and four months at the time of the separation This seems to be rather convincing evidence of a decided environmental influence. The chief weakness of the study lies in the fact that all the environmental influences have not been eliminated, since these siblings had for a period of years been subjected to a common environment. The more than five years spent in the same home probably accounted for a part of the correlation found. In fact, present day tendencies in psychological theory would seem to indicate that the most significant period of their life had been spent in a similar environment From the data which they had at hand it was impossible to determine the extent of this influence.

The study reported in this paper represents a different attack on the same general problem The procedure used here in the attempt to answer the question of the relative influence upon intelligence of blood relationship and common environment was to compare the correlation between the intelligence of pairs of siblings from the same home with the correlation between pairs of unrelated children, the unrelated pairs being equated with the sibling pairs on the basis of age, school attended, and home background The significance of the correlation between the intelligence of siblings can only be interpreted in the light of information as to the correlation that would be found if the members of the pairs were unrelated Presumably children paired at random would show no resemblance, but one cannot compare paired siblings with random pairs and account for the differences in terms of inheritance. It is only when environmental influences are equal that comparisons have meaning. In this study the attempt has been made to equate two sets of paired children on the basis of environment, the members of each pair in one set having the same parents while in the other the members of a pair have different parents. To the extent that these conditions have been met, the difference in the degree of relationship between the members of the first set compared with the second set can be considered as a measure of the influence of common parentage.

During the spring of 1927 the writer cooperated in the testing of all children found in Grades V to XI in five school systems in Lincoln Parish, Louisiana Four of these schools were eleven-grade consolidated rural schools, the fifth system consisted of two elementary schools and a high school located in a town of approximately 5000 population. The Otis Self-administering Tests of Mental Ability were used to measure intelligence, and the Sims Score Card for Socio-economic Status was used as a measure of home background The Otis Intermediate Test, Form A, was used for Grades V to VII, and the Higher Test, Form A, for Grades VIII to XI. On the basis of the scores made on the Otis Tests the IQ's were determined by the procedure recommended by Otis in his "Manual of Directions." The reliability of the Otis Tests, as reported by the author, is high (.95 for the Intermediate Form and .92 for the Higher Form). The validity is more difficult to determine, but high correlations reported between this test and other tests of intelligence would seem to indicate that it is measuring approximately the same thing that all of our intelligence tests measure The reliability and validity of the Sims Score Card has been discussed at length elsewhere.[1] Suffice it to say here that it has been shown to adequately differentiate between groups with known differences in home environment; and the reliability, as determined by correlating the scores of children from the same home, has been found to be rather high, approximately 90. The coefficient of reliability determined from one hundred pairs of sixth-, seventh-, and eighth-grade children was 95, and the coefficient found by correlating the scores of the two hundred three pairs of siblings reported below was .87. The relatively low reliability of the data used here may be due to the fact that some of the examiners were not very skilled in testing, but there is evidence that high school children report a slightly higher socio-economic status, perhaps because certain of the items included in the scale are actually more frequently possessed by older children This increase is, however, very slight, being less than one-third of a point a year, and, although it does lower the reliability of the measure, it is still reliable enough for our purposes.

From the tested population described above, each child reporting a brother or sister living in the same home and enrolled in one of the schools and grades included in the testing (Grades V to XI) was selected and his record paired with that of the brother or sister reported. These siblings were paired in all possible ways, that is, where three siblings were tested they were grouped *A* with *B*, *A* with *C*, and *B* with *C*. More than three siblings were not reported in any case. This pairing seems to be the generally accepted one when the number of siblings found in any one family is not large Two hundred twenty-four such pairs, coming from one hundred eighty-two homes, were secured, but for reasons explained later, twenty-one of these pairs were discarded, so that two hundred three pairs from one hundred sixty-one homes composed the main group used in this study.

---

[1] "The Measurement of Socio-economic Status." Public School Publishing Co, 1928

The unrelated pairs were prepared by substituting for one member of each sibling pair an unrelated child coming from the same school, and having the same age and home background. The procedure for preparing one of these unrelated pairs was as follows. Referring to the school where the younger sibling was enrolled, that child with the same home background score and most nearly the same age as the younger sibling was selected. When two or more children were found to fulfill these requirements to the same extent, one was chosen at random from the number. In the same manner the child most nearly a duplicate of the older sibling was selected The one of these two selected children nearer the age of the sibling which he was selected to duplicate was taken as one member of the unrelated pair and the other sibling as the second member. In no case was a child selected whose age was not within one year of the sibling and whose home background score was not within one point of the sibling's score, or the average of the siblings' scores in cases where there was a difference in the reported home background score of two siblings If no such child could be found, the sibling pair was discarded. The twenty-one pairs of siblings mentioned above as being eliminated from the study were discarded because no unrelated pair could be found to match them

In this manner a group of two hundred three pairs of unrelated children having the same age and home background and coming from the same school as the sibling pairs were secured Within the reliability of the measures used they differ from the sibling pairs only in the absence of common parentage. Table I presents the average age and socio-economic status (with the standard deviation) of the two groups, the young members of the pairs being contrasted with the old members In addition, although it is to be noted that they have not been paired on these bases, the mean grade and IQ (with the standard deviation) is reported.

That the groups used here are fairly representative of the population from which they were selected is shown by comparing the averages and sigmas for intelligence and socio-economic status of the siblings and unrelated pairs with the total population. The total population tested consisted of 1018 cases and the average IQ was 90.9 (with a standard deviation of 15.6), while the average socio-economic status was 14.7 (with a standard deviation of 5.9)

TABLE I.—SHOWING THE MEAN AGE, SOCIO-ECONOMIC STATUS, GRADE AND IQ OF THE SIBLING PAIRS AND THE UNRELATED PAIRS

	N	Age		S-E-S		Grade		IQ	
		Mean	Sigma	Mean	Sigma	Mean	Sigma	Mean	Sigma
Young member									
Sibling	203	13.0	2.0	12.4	5.4	7.2	2.1	91.8	15.4
Unrelated	203	13.0	1.9	12.9	5.4	7.3	1.7	91.3	14.7
Old member									
Sibling	203	15.9	2.1	13.4	5.6	9.1	1.8	87.3	13.6
Unrelated	203	15.9	2.2	13.5	5.4	9.1	1.8	87.5	13.3
Difference between old and young									
Sibling		2.9		1.0		1.9		4.5	
Unrelated		2.9		.6		1.8		3.8	

Comparison of the sibling and unrelated groups shows their great similarity. The differences between the members of the sibling pairs (young against old) is seen to be practically the same as the differences between the members of the unrelated pairs The average difference in age is approximately three years, and in grade approximately two grades; the old members score slightly higher on socio-economic status; and the younger members have the higher IQ One is struck by the fact that, although no attempt has been made to equate the groups on the basis of either grade or intelligence, there appears to be just as much similarity between the two groups on these measures as on the other two

Certain factors may independently affect the intelligence of members of pairs such as those here used; consequently the degree of correlation found between the paired cases is influenced by the method used for entry in the correlation table. Because there is yet some doubt as to what is the most satisfactory method, in the correlations reported here two methods were used. By the first (known hereafter as the age-pairing method) the score of the young member was always entered on the vertical axis of the correlation table and the score for the corresponding old member on the horizontal axis. By the second (known as the double-entry method) each pair of scores was entered twice, with the young on the vertical axis and then with the old on the vertical axis.[1]

The correlation between the IQ's of the siblings and the correlation between the IQ's of the unrelated pairs, using the two methods, are as follows:

	r (Age Pairing)	r (Double Entry)
Siblings ($N = 203$)	44 ± 04	40 ± 04
Unrelated ($N = 203$)	35 ± 04	29 ± 04

On the surface, at least, it seems that these correlations, between the intelligence of the members of the sibling pairs and between the intelligence of the unrelated pairs, should be indications of the relative influence of *common environment plus common parentage*, and of *common environment only* on intelligence as measured by a group test, that is, common environment produces a correlation of 35 or .29 depending upon the method used, while the addition of common parentage raises this correlation to .44 or .40 again depending upon the method used The writer is inclined to believe that these are the most significant correlations reported, but there are undoubtedly certain factors which might influence these coefficients that make it undesirable to accept them at face value. It has been pointed out by numerous investigators that there is a decided relationship between age and intelligence, consequently the coefficients for the members of each group is being thus affected. For comparative purposes, since the age factor is presumably operating the same in both cases, it is perhaps of small matter; but it seemed desirable to see what the correlation would become when age is kept constant. Partial correlations with age constant were determined for the sibling pairs and for the unrelated pairs. In determining these correlations the age pairing method was used. These correlations were, for the sibling pairs 48, and for the unrelated pairs .34 In other words, this correction for age has not seriously affected the relation between the coefficients for siblings and for unrelated pairs

---

[1] In using these two methods we are following the procedure of Freeman, Holzinger, and Mitchell in the study mentioned above

It will be recalled that two forms of the Otis tests were used for testing the intelligence of the subjects Otis reports a correlation of only 84 between the two forms, consequently IQ's determined from different forms are not exactly comparable. In order to determine the influence which the use of two forms might be having on the correlations found, those pairs where the child substituted to make an unrelated pair was not tested with the same form as was the sibling for which he was substituted were first eliminated[1] Coefficients of correlation were then determined for: (1) those cases where both members of the pair were tested with the intermediate form, (2) those cases where both members of the pair were tested with the higher form; (3) those cases where the young member was tested with the intermediate form and the old member with the higher form, and (4) the composite of all cases. The correlations between the intelligence of siblings are compared with those between the unrelated pairs in Table II Considering the composite of these cases it will be

TABLE II—COMPARISON OF THE CORRELATION BETWEEN INTELLIGENCE OF SIBLINGS WITH UNRELATED WHEN THE SAME FORM WAS USED

	r (AGE PAIRING)	r (DOUBLE ENTRY)
Intermediate form		
Siblings ($N = 64$)	49 ± 06	49 ± 06
Unrelated ($N = 64$)	37 ± 07	32 ± 08
Higher form:		
Siblings ($N = 55$)	55 ± 06	48 ± 07
Unrelated ($N = 55$)	43 ± 07	36 ± 08
Both forms:		
Siblings ($N = 56$)	45 ± 07	41 ± 07
Unrelated ($N = 56$)	31 ± 08	31 ± 08
Composite,		
Siblings ($N = 175$)	51 ± 04	49 ± 04
Unrelated ($N = 175$)	30 ± 04	37 ± 04

---

[1] Since an unrelated pair was made by substituting that child in the school who had most nearly the same home background and age, but not necessarily the *same grade*, as the sibling, it sometimes happened that the substituted child had been tested with a different form from that used on the sibling

noted that although the correlations have been raised slightly there is no serious change in the ratio of the unrelated to the sibling. Comparing the correlation found when the members of a pair were tested by the same form with that found when the members of a pair were tested one by one form, and the second by the other, we see evidence that the use of two forms is probably reducing the correlation, but we also see that the siblings and unrelated are reduced to the same extent.

A final factor that needs to be taken into consideration in the interpretation of the coefficients that have been found in this study is that of selection. What would be the effect on the correlation found if we had available all of the brothers and sisters of the siblings used? What would be the effect if we had all of the children in the community, or even in these schools, who were of an age and from such homes as to make them suitable for use in the preparation of our unrelated pairs? We have made an attempt to answer these questions. Presumably pairs with age differences such as those of the cases used here made up at random would show no correlation between intelligence. As a check on the validity of the method of pairing used here—in order to see if it operated in such a way as to select pairs with like intelligence independent of the factors which we have assumed to be causing the similarity—a third set of paired cases was prepared. These paired cases were made by selecting for each sibling that child, in the same school (and in the grades included in the study) having most nearly the same age as the sibling. When the nearest age was found to be represented by several children, one was selected at random. A duplicate for each sibling pair was then secured by taking the selected child nearer in age to the sibling he was selected to duplicate as one member, and the other sibling as the second member. In this manner a set of two hundred three pairs of unrelated children having the same age and coming from the same school as the siblings was prepared. Since they have been paired without regard to either blood relationship or like home background, the correlation should be zero. The correlation actually found, by age and by double entry-pairing respectively, was .05 ± .05, and .04 ± .05. It appears that whatever selection is operating is very slight, and it is presumably affecting the siblings and the unrelated in the same way.

To summarize, then, siblings reared in the same home have been found to have a correlation of approximately .45 between their intelligence quotients as determined by a group test, while unrelated pairs equated with the sibling pairs on the basis of age and home background, and attending the same school as the siblings, have shown a correlation between their IQ's of something in the neighborhood of 30.

Attention should be called to the fact that the crudeness of our measure of home background is operating in such a way as to increase the difference between these two correlations The siblings have been reared in the *same home*, and to the extent that this means common environment we are sure they have it The unrelated pairs have been reared in different homes, but homes we have assumed to be alike because they had the same score on the score card. To assume that two children have the same home background because they have the same score on this scale is unreliable to the extent that the measure is unreliable, but to compare children equated on this basis with children actually reared in the same home has an added factor of unreliability caused by the failure of the score card to measure all of the finer environmental differences found among homes that are seemingly equal The coefficient for the siblings is not subjected to this factor, which is almost sure to attenuate the correlation for the unrelated pairs. In other words, if we could find unrelated pairs that had actually been reared from birth in the same home we would expect a correlation between unrelated pairs that would more nearly approach that found for siblings One feels sure that the influence of blood relationship on the intelligence of siblings is not being underestimated when the difference between the coefficient found here for the siblings and for the unrelated pairs is taken as its measure. Since unrelated children who, within the reliability of our measures, have been equated with siblings show the resemblance that we have found, it seems safe to say that the common environment to which the siblings are subjected accounts for a correlation between their IQ's of at least .30, while blood relationship is potent enough to raise this coefficient to .45

In conclusion we should call attention to the fact that throughout this paper we have been considering intelligence as measured by a group test The fact that high correlations have been found between various group tests of intelligence, and between group tests and individual tests, would lead one to believe that similar results would be found whatever the test used This, however, does not warrant the assumption that there is not some quality in the human, some native capacity, which is inherited. The results may be interpreted as a condemnation of present day 'intelligence' tests or as evidence that intelligence is not solely inherited, but rather a development due to the interplay of environmental forces on hereditary characteristics. The interpretation that one may give the findings is a personal matter, but the indications are that intelligence, so far as we are today able to measure it, is greatly influenced by environment.

THE EQUIVALENCE OF JUDGMENTS TO TEST ITEMS IN THE SENSE OF THE SPEARMAN-BROWN FORMULA[1]

H. H REMMERS

Purdue University

THE PROBLEM

A previous study by two of my students and myself[2] reported an investigation in which a summary of the experimental literature indicated that "the Spearman-Brown prediction formula shows it to give meaningful prediction on such materials as mental test items, spelling words, [judgments of] lifted weights, true-false items in language, and component units of rating scales." This previous work was done by a number of different authors—Kelley, Holzinger, Clayton, Ruch, Ackerson, Jackson, and Furfey.

Our own experiment at that time reported results on the Purdue Personnel Rating Scale which indicated that summations of ratings on ten different traits fell within the limits of allowable error when the values of empirically observed correlations for varying numbers of raters were compared with those predicted by the Spearman-Brown formula That is, the formula did predict with reasonable accuracy the reliabilities to be expected with an increase in the number of judges or raters

The typical situations in which judgments or ratings are obtained is frequently limited by the fact that, assuming the Spearman-Brown formula to apply, the number of judges is too small to give ratings sufficiently reliable for practical purposes

The present study is the report of an investigation of the Purdue Rating Scale for Instructors in which the limitation just mentioned is largely absent, since the average instructor is likely to have a hundred or more students On this scale students judge instructors anonymously by checking on a graphic scale the amount of ten different traits presumably related to success in classroom teaching possessed by a given teacher under consideration. In a rating program carried out at Purdue University during 1928–1929, something over ninety per cent of the faculty were rated [1] A number of instructors turned over to me their marked and scored rating blanks The question the answer to which was sought was, Do the judgments which students record concerning their instructors follow the law represented by the Spearman-Brown prophecy formula? In other words, is it valid to assume that judgments are the equivalent of test items in the sense required by the formula?

---

[1] A paper given before Section I of the American Association for the Advancement of Science, Dec 28, 1929, Des Moines, Ia.

[2] Remmers, H H., Shock, N. W., and Kelly, E. L An Empirical Study of the Validity of the Spearman-Brown Prophecy Formula as Applied to the Purdue Rating Scale *Journal of Educational Psychology*, Vol XVIII, No. 3, March, 1927, pp. 187–195

The formula, by now rather well known in its application to test construction, is

$$r_n = \frac{nr_{11}}{1 + (n - 1)r_{11}},$$

where r_n = the predicted reliability,
n = the number of times a test is increased by its own length, and
r_{11} = the reliability of a unit length of the given test.

In the present problem the unit is the judgment of a single student, and the problem is to determine the correspondence of reliability coefficients determined for given numbers of such judgments with those predicted by the formula.

If the conditions of the formula were always fully met, perfect prediction would result Practically, however, the conditions are never ideally met, and our problem becomes a sampling problem subject to the laws of probability

It should be recalled that the validity of the formula depends upon the homogeneity of the test items, or, statistically speaking, upon the equality of intercorrelations between units and upon the equality of the standard deviations of these units

PROCEDURE

In the ratings of the instructors used for this investigation there was, so far as I know, no selective factor operating in the return of these blanks, except that instructors with unusually low ratings might have hesitated to return them as readily as those with more satisfactory ratings To the extent that such a factor did operate it would tend to reduce the variability of the total distribution of ratings and thus to reduce the reliabilities below what they would otherwise have been

---

[1] Remmers, H H : The College Professor as the Student Sees Him *Bulletin* of Purdue University, Studies in Higher Education, Vol XI, March, 1929, p 63

From each instructor's ratings scores were selected according to the following schema:

20 samplings of 1 *vs* 1 rating for trait 5, Presentation of Subject-matter
5 samplings of Σ 5 ratings *vs.* Σ 5 ratings, for trait 5, Presentation of Subject-matter
5 samplings of Σ10 ratings *vs* Σ10 ratings, for trait 5, Presentation of Subject-matter
5 samplings of Σ15 ratings *vs.* Σ15 ratings, for trait 5, Presentation of Subject-matter
1 sampling of Σ20 ratings *vs* Σ20 ratings, for trait 5, Presentation of Subject-matter
1 sampling of Σ30 ratings *vs* Σ30 ratings, for trait 5, Presentation of Subject-matter

Traits 1 and 10, Interest in Subject-matter and Stimulating Intellectual Curiosity, and also the sums of all traits, were sampled in exactly the same way, except that only ten 1 *vs.* 1 correlations were calculated for these, and that in the case of the sums of all traits the samplings obtained included only 1 *vs* 1, Σ20 *vs.* Σ20, and Σ30 *vs.* Σ30.

The results of the samplings are shown in Tables I to VII.

THE DATA

TABLE I—DISTRIBUTIONS OF CORRELATIONS 1 *vs.* 1 RATING FOR THE TRAITS INDICATED

	Trait 1	Trait 5	Trait 10	Sums of all traits
.60 to 69		4		2
50 to 59	1	3	2	3
40 to 49	1	5	3	2
30 to 39	5	3	2	1
20 to 29		3	1	
10 to 19	1	2	1	1
00 to 09	1		1	
− 10 to − 01	1			
− 20 to − 11				1
Total	10	20	10	10
Mean	290	429	354	320
Median	344	450	363	503
N*	37	37	37	37

* N in all tables refers to the number of instructors

In Table VII are summarized the necessary data for answering the problem raised at the outset of this study, *i.e.*, as to whether judgments under the defined conditions are equivalent to test items in the sense of the Spearman-Brown formula The answer is that they are. There is apparently a slight tendency for the formula to overpredict, there being sixteen overpredictions out of a possible seventeen. These apparent overpredictions, however, are probably to be explained by the reduction in the number of instructors necessitated by the lack of instructors with very large classes This tended to reduce the "range of talent," with a corresponding reduction in reliability. In no single

Table II —Correlations of $\Sigma 5$ *vs* $\Sigma 5$ Ratings for the Traits Indicated

	Trait 1	Trait 5	Trait 10
	863	783	615
	799	782	576
	752	760	433
	721	709	428
	705	644	345
Mean	728	736	479
Median	752	760	433
N	36	36	36

Table III.—Correlations of $\Sigma 10$ *vs* $\Sigma 10$ Ratings for the Traits Indicated

	Trait 1	Trait 5	Trait 10
	843	890	861
	.832	856	781
	663	.856	754
	652	843	661
	596	782	530
Mean	.717	845	.717
Median	663	856	.754
N	20	20	20

Table IV —Correlations of $\Sigma 15$ *vs* $\Sigma 15$ Ratings for the Traits Indicated

	Trait 1	Trait 5	Trait 10
	961	973	902
	882	954	876
	702	931	872
	621	904	821
	607	571	791
Mean	751	887	852
Median	702	931	872
N	7	7	7

TABLE V —CORRELATIONS OF Σ20 *vs.* Σ20 RATINGS FOR THE TRAITS INDICATED

	Trait 1	Trait 5	Trait 10	Sum of all traits
	.801	877	.636	898
N	13	13	13	40

TABLE VI.—CORRELATIONS OF Σ30 *vs.* Σ30 RATINGS FOR THE TRAITS INDICATED

	Trait 1	Trait 5	Trait 10	Sum of all traits
	936	876	805	872
N	10	10	10	16

case, however, is the difference between the "best" or most probable of the observed correlations and that of the corresponding predicted values clearly statistically significant In the case of only two correlations is the difference divided by its probable error greater than two

On the face of the data, it seems surprising that the summation of all traits should give no higher reliability than was observed This, however, is probably to be explained by the fact that the raw trait scores are incommensurable Had these scores been first reduced to standard measures or percentile equivalents, the resulting reliabilities would very probably have been increased very materially.

SUMMARY AND CONCLUSIONS

Samplings of judgments for varying numbers of judges selected at random yielded reliability correlations within the allowable error when the judges are undergraduate students and the things judged are classroom personality traits of instructors who are appreciably different in that they possess, in the judgments of students, different amounts of these traits. The number of correlations upon which this study was based is one hundred three The following generalizations seem warranted from these and previously reported data

1 Reliabilities are predicted within the allowable error up to thirty judgments

2. The three traits sampled in this investigation vary significantly in reliability Stimulation of Intellectual Curiosity, for example, means more different things to students than does the trait Presentation of Subject-matter.

3 In general, ratings by from ten to twenty students on a single trait for instructors differing sensibly in the amount of the trait pos-

TABLE VII —CORRELATIONS OBSERVED AND PREDICTED

	No r's computed	No of instructors	Average observed	Correlation predicted[1]	Diff	PE$_{diff}$	$\frac{\text{Diff}}{\text{PE}_{diff}}$
Trait 5 Presentation of subject-matter							
1 *vs* 1	20	37	410 ± 001				
Σ 5 *vs* Σ 5	5	30	736 ± 052	783 ± 001	047	082	,6
Σ10 *vs* Σ10	5	20	845 ± 013	878 ± 051	033	,060	5
Σ15 *vs* Σ15	5	7	887 ± 053	015 ± 007	028	085	3
Σ20 *vs* Σ20	1	13	877 ± 013	035 ± 038	058	057	1 0
Σ30 *vs* Σ30	1	10	876 ± 050	056 ± 030	080	058	1 4
Trait 10 Stimulating intellectual curiosity							
1 *vs* 1	10	37	351 ± 007				
Σ 5 *vs*, Σ 5	5	30	470 ± 086	733 ± 080	254	122	2 1
Σ10 *vs* Σ10	5	20	717 ± 073	810 ± 077	120	106	1 2
Σ15 *vs* Σ15	5	7	852 ± 070	802 ± 006	010	110	3
Σ20 *vs* Σ20	1	13	030 ± 111	015 ± 055	270	121	2 3
Σ30 *vs* Σ30	1	10	805 ± 075	013 ± 015	138	087	1 6
Trait 1 Interest in subject-matter							
1 *vs* 1	10	37	200 ± 102				
Σ 5 *vs* Σ 5	5	30	728 ± 053	671 ± 110	057	122	5
Σ10 *vs* Σ10	5	20	717 ± 073	803 ± 100	080	120	7
Σ15 *vs* Σ15	5	7	751 ± 110	800 ± 137	106	176	6
Σ20 *vs* Σ20	1	13	805 ± 000	801 ± 081	087	105	8
Σ30 *vs* Σ30	1	10	030 ± 026	025 ± 060	,011	071	2
Sums of all traits							
1 *vs* 1	10	37	320 ± 006				
Σ20 *vs* Σ20	1	40	808 ± 021	004 ± 038	000	043	1
Σ30 *vs* Σ30	1	16	872 ± 010	933 ± 043	061	059	1,0

sessed yield reliabilities which compare rather favorably with the reliabilities reported for standardized mental and educational tests

4. It is probable that in the majority of situations in which subjective judgments are used—personnel ratings, stock judging, debate judging, beauty contests, jury verdicts, political polls, etc.—the Spearman-Brown prophecy formula indicates the number of judgments required for a given reliability, although here it must be admitted that we are going beyond the known facts

---

[1] These r's were calculated by means of Shen's formula, $PE_R = \frac{0745a(1 - r^2)}{\sqrt{N}[1 + (a - 1)r]^2}$ See his article, A Note on the Standard Error of the Spearman-Brown Formula *Journal of Educational Psychology*, Vol XVII, 1026, pp 03–04, also Douglas, Harl A , A Note on the Correctness of Certain Error Formulas *Journal of Educational Psychology*, Vol XX, 1020, pp 434–437

SEMI-LOGARITHMIC VERSUS LINEAR PLOTTING OF LEARNING CURVES

RICHARD W. HUSBAND

University of Wisconsin

The usual methods of plotting learning curves, say of time or of errors, using linear functions in both variables, have certain limitations

1. The actual shape of the curve will vary considerably with different spatial separation of successive units If the spread is wide, the whole curve will be rounded out, approaching the so-called "typical concavity." If it is small, the curve will be flattened out near the base line

2. The unit gains tend to be artificial, and dependent on the arbitrary intervals chosen. Reduction from ten to nine errors is not nearly as important or as difficult to accomplish as from two to one or from one to zero, yet on a linear diagram the two appear equal

3 A further difficulty lies in making objective and decisive comparisons between curves Aside from theoretical and technical interest, the chief functions of quantitative experimentation, including the subsequent statistical treatment, are to compare the relative performances or efficiencies of different groups or techniques Examples are such questions as two different spacings of learning periods, dietary or drug conditions, or comparisons between several species.

Just as Thorndike and Koffka interpret stupidity and insight respectively from the same curves, so may two investigators read clear superiority and negligible differences from a single set of comparative curves. The same ambiguity is present in comparing means unless we use probability figures derived from the standard error of the difference between the means. With linear plotting we do not have available any *single* figure to use in expressing entirely apart from personal opinion the differences between performances in learning, especially considering that we are dealing with a continuous temporal series. We are more or less limited to admiring the curves from an aesthetic standpoint, and passing a few judgments from inspection

Occasionally an ambitious investigator has attempted to fit the curves he has obtained to an equation, in order to express the progress of learning in a single figure or in a single function This is all very well, but such equations are usually highly specialized, and are only suitable for the single set of data. They often involve advanced mathematics, not an available tool to all experimenters; use logarithms, trigonometry, or calculus, and difficult-to-determine constants of only individual application.

All these difficulties melt away at once if one plots the same data on semi-logarithmic paper. The horizontal axis or abscissa is used for successive trials, as usual, and is kept on a linear basis, as each trial is by nature equally separated from all others. The vertical axis, or ordinate, is scaled logarithmically, and the score under consideration, errors, time, or units done, as the case may be, is plotted on this. In some cases, owing to the manner in which the paper is printed, the original scores will have to be converted into percentages, say of initial or of final performance. This, however, is no serious handicap, as the original unit scores may be written on the axis as well. No knowledge of logarithms is needed; plotting is done as directly as when linear paper is used. The graph paper itself takes care of this function.

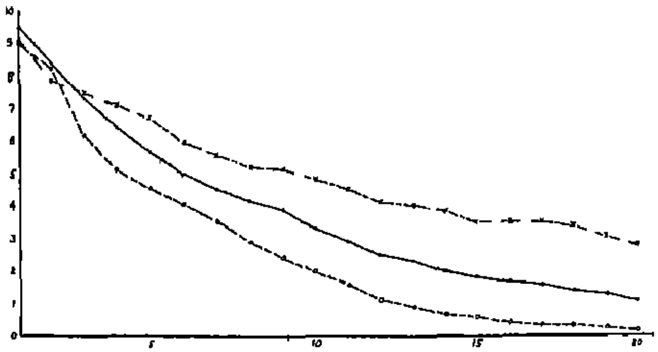


CHART I.—Linear learning curves. (Ordinate) average number of errors. (Abscissa) trial.

This procedure disposes of the difficulties we have raised as follows: As to the first, the shape of the curve, there will be no difference, no matter how it is plotted. Regularly progressive learning will invariably result in a straight line. The second objection is disposed of, as the fundamental principle of logarithms is proportion rather than absolutes. On this basis equal proportional gains are represented equally, no matter at what point in the learning process. Thus, reduction from eight to four errors does not show up as any greater than does later improvement from four to two, or from two to one. Therefore, if an individual learns a constant fraction of the material on each trial the successive points will make a straight line when connected together. This leads us directly to solution of the third defect of the former method, that of objective comparisons between curves If the two or more curves under consideration take practically straight lines, we can compare progress made by the various groups by quoting

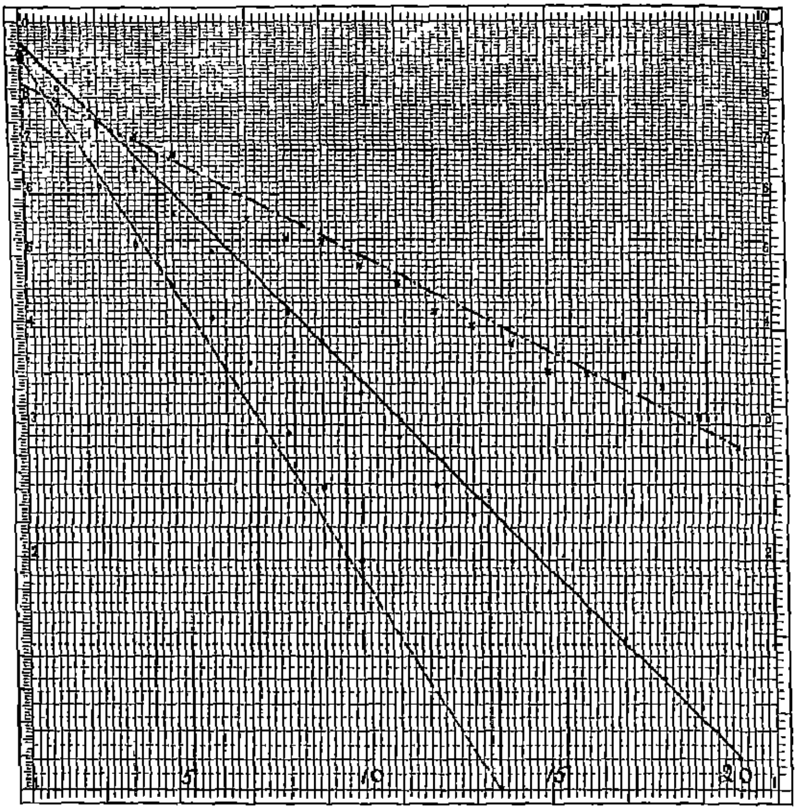


CHART II —Semi-logarithmic learning curves (Ordinate) average number of errors. (Abscissa) trial

the number of trials required by each to reduce by half the error or time score, or to double the output.

We include learning curves drawn from the same data by the two methods. The data were obtained by the writer from a group of eighty subjects who learned a maze of moderate difficulty. In both cases the solid line represents the figures for the full group (eighty), the broken line those (twenty-seven) in the group who learned by a purely ideational method, and the line made up of a dash and two dots the motor learners (fifteen). The data were smoothed by a moving average of three trials, in order to reduce minor fluctuations. This in no way makes any difference in the comparisons, since both sets of curves were treated alike The straight lines drawn in the logarithmic data are only approximate, it must be admitted, but the deviations of the individual points are seem to be remarkably small.

Let us now give sample interpretations of the two sets of curves.

(*A*) *Linear Method.*—(1) All groups start at about the same level. (2) The motor learners soon are at a disadvantage, while those using an ideational attack do better than average (3) The differences appear to widen up to about the tenth trial, after which there is doubt

(*B*) *Logarithmic Method*—(1) and (2) the same as under (*A*) (3) The absolute differences, as well as relative, are constantly widening, and as far as the twentieth trial there is no tendency to narrow down. (4) Learning, in this problem at least, is characterized by a proportionate reduction of errors, which shows up in this chart as a straight line (5) Judging by reduction of errors from six to three (or any other arbitrary limits) the average individual halves his inaccuracies in 6½ trials, the ideational learner in 4½, and the motor learner in fourteen attempts

Do not these last figures alone mean much more than any amount of verbal discussion, description, and estimate? In other words, we lose nothing but gain a great deal. Yet no complicated mathematics is needed in order to make one's interpretations and comparisons. All in all, plotting data on a logarithmic basis gives us clearly objective and concise ways of interpreting data

The writer does not claim that all learning curves will form a straight line if plotted this way But in trying sixteen different sets of figures from maze learning records, there was not one in which the points did not hover within the normal variability limits of a frequency on either side of the lines which were drawn. Slight curvilinearity would not stand in the way of interpretations on the basis we have described.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

CONDUCTED BY FRANCES M. FOSTER

Fundamentals of Educational Psychology, by I M. Gast and H. C Skinner New York. Benj H Sanborn and Co , 1929. Pp. XIII + 354

Changing viewpoints in educational psychology necessitate fresh treatments of the subject The aim of the authors of the present volume has been to relate the subject more closely to classroom practice, to include the social and applied aspects, and to present the best of new opinion and evidence of new movements in education and psychology. Their account is expressed in terminology that the average student can understand, uncomplicated by highly technical details. Educational applications are made for every topic discussed. The most recent reliably determined evidence in support of statements is consistently referred to The authors state that they have not attempted to present a "logical system of psychology" They have gleaned from a wide variety of sources and from varied schools of psychology explanations of phenomena which appear to them to be most authentic. The result is a presentation of the subject that contains much of the older material, the more familiar topics one is accustomed to seeing in a standard text, together with new points of view and an occasional new chapter, "Mental Health" "Childhood and Adolescence," "The Gifted Child," "Nature and Nurture," "Intelligence—Its Nature and Measurement," "Educational Measurements." A list of the psychologists who have contributed most to education is given, and a unique feature of the book is the inclusion of their photographs Questions and problems for further study, a list of references cited in the text, and suggested readings follow each chapter

The authors have avoided critical discussion of divergent opinions and antagonistic theories of different schools of psychology The newer interpretations of gestalt psychology are scarcely mentioned Description of the physiological bases of behavior has been limited to a few of the most essential facts which are stated in the introductory chapter The authors appear to be committed to a "conditioned reflex" interpretation of learning, although the discussion is not at all times consistent. Their explanations of learning, habit formation, and skill impresses the reader as being the least commendable of all the topics discussed The educational applications contain many excellent ideas which might well have been expanded or more fully explained Too often they are in the form of advice which is not closely related to the psychological principles presented and not pertinent to the subject discussed The result is occasional lack of coherence These defects are minor and do not detract from the usefulness of the book as a source of material for the student of educational psychology. GERTRUDE HILDRETH.

The Lincoln School of Teachers College.

Studies in The Organization of Character, by Hugh Hartshorne, Mark A May, and Frank K. Shuttleworth New York. The Macmillan Co Pp. 503

This volume brings to a close the work of the "Character Education Inquiry." It follows "Studies in Deceit" and "Studies in Service and Self-control," and is in many respects the most important volume of the three. It contains the following parts· I. Social Intelligence and Social Attitude; II Interrelations of the Factors of Character; III. Components of Character; IV. The Significance of Integration; V Conclusions of the Character Education Inquiry.

Those who have followed the work of the "Character Education Inquiry" will recall that the investigation showed that deception, helpfulness, cooperation, persistence, and inhibition were groups of specific habits, rather than general traits. "One could hardly predict from what a person did in one situation what he would do in a different situation" There are, however, obvious differences in individuals in the degree to which their behavior is of a consistent pattern. It is chiefly with the study of this inner consistency that this volume is concerned

To eight hundred fifty children in the fifth to eighth grades of three school groups (or composites) were given a large battery of tests, including tests of moral knowledge and attitudes, tests of deception, of cooperation, of inhibition, of persistence, of suggestibility, of "nervous instability," and various rating schemes, for determining "level of social functioning," reputation, and the like Measures of intelligence and home background were also available

The correlations between the various objective tests, even when corrected for attenuation, showed little relationship between gross trait scores. For example, honesty correlates with service 439, with inhibition .487, with persistence .166, "This meager consistency in behavior of good or socially desirable types may seem lamentable, but we shall have to conclude that consistency is not characteristic of children . . We cannot infer from the conduct tests the presence of a general factor."

The effect of group standards on conduct was investigated, and while there was little relationship between an individual's ethical understanding and his conduct (aside from the relationship existing through intelligence), there were definite relationships (aside from intelligence) between knowledges of right and wrong and actual conduct when classroom groups were considered as a whole

Probably the most immediately valuable part of this report is that dealing in a very detailed manner with the integration of honesty. Other of the character traits might have been investigated, but the honesty tests were used since they were the most comprehensive battery available It will be remembered that the intercorrelation between the various honesty tests was very low, leading necessarily to the doctrine of specificity of reaction in this field. Now if the scores of a child in each test be distributed, and a measure of dispersion computed, we have a measure of the degree to which the child tends to be consistent in his behavior—be it honest or dishonest Twenty-one tests were used For each child the SD of his scores was computed This is called his index of integration Thus a perfectly honest (or dishonest, or 50 per cent honest) child would have an index of 0.00. When the unreliability of the tests used was compensated for, there still remain large differences in integration These integration indices were correlated with total honesty score The correlation, corrected for attenuation, is about 80. This means that the more honest a child is the more consistent is he in his honesty. The possibility that this result is an artefact of the nature of the testing set-up is interestingly and fairly discussed, pro and con. It is concluded that in so far as the test situation is responsible for the correlation, it is not artificial in that the same factors which make the more honest children more consistent in their test results also are operative in the life situations calling for honesty

The implications of the above are interesting Although moral conduct is, in the main, largely specific, it is shown that as conduct is improved there results a general improvement It is the behavior of the dishonest child which can be most readily modified

It is difficult adequately to appraise a study so vast in its scope as this work It is certain, however, that this Inquiry is of really tremendous importance to education. And the present volume is easily the most interesting and valuable to the non-research person in the educational field. DONALD SNEDDEN

New York University.

Research Methods and Teachers' Problems, by Douglas Waples and Ralph W. Tyler. A Manual for Systematic Studies of Classroom Procedure. New York. The Macmillan Co, 1930 Pp. XXIII + 653.

The solution of administrative and instructional problems by research methods is an important aspect of modern educational practice The research worker now has at his disposal a group of publications describing research technique to which the present book, designed as a text for teachers, administrators and technicians is an important addition. Research is interpreted by the authors' in a broad sense as including not only laboratory experimentation but the study of service problems such as the teacher or administrator meets in operating a school. The authors, however, distinguish between these two types of research, pointing out the narrower scope and more intensive approach in laboratory studies Such a problem as testing and classifying a group of pupils is considered research though distinguished from laboratory experiments in measurement and classification

The authors have devised a working plan for research projects in general, divided into several major steps The types of data needed for typical problems are listed, available sources of data are enumerated, and methods of obtaining data from available sources are outlined The research problems commonly met by educators are classified in three groups, problems of the curriculum, of teaching methods, and of school management. For each of these three classifications appropriate methods of attack are outlined and illustrative studies are presented in outline form A list of sixteen different techniques or methods of approach with an evaluation and description of the method of applying each concludes the book Adequate bibliographies are provided at the close of major topics

No book on research methods can of itself produce experts in educational research, just as the expert cook is not necessarily the product of a cook book, but just as the cook's technique improves with training in method, so may the research student improve his experiments with increased training in technique. To follow any treatise on research methods slavishly would be to defeat the very purpose of research. In using the book under discussion some students may be inclined to accept too readily the interpretations of the authors and the carefully detailed outlines without exercising originality or independence of judgment. When the outlined pattern fails to fit, the student must be resourceful and intelligent enough to alter or adapt it. The authors unfortunately give no discussion of the use of research in modernizing educational methods. GERTRUDE HILDRETH.

The Lincoln School of Teachers College.

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII February, 1931 Number 2

PRACTICE VERSUS GRAMMAR IN THE LEARNING OF CORRECT ENGLISH USAGE[1]

PERCIVAL M SYMONDS

Teachers College, Columbia University

Recent correlation studies have shown little relationship between knowledge of English grammar on the one hand, and the correctness of usage on the other Those who are familiar with these studies commonly believe that the value of English grammar in influencing correct grammatical usage in speaking or writing has been definitely discredited. But correlation studies have frequently misled in the past regarding the actual dynamic relationships in learning and the amount of transfer that can be expected. Grammar still has a somewhat thriving existence in our schools

In 1906 Hoyt[2] reported a correlation as low as 18 between grammar and composition ability among high school pupils. Rapeer[3] reports the same relationship to be 23. Segal and Barr[4] in a minor study reported a correlation of 48 between a test of formal grammar and a test of "applied grammar" which is really a test of ability to choose the correct form. Asker[5] reached similar conclusions by a different method. The general conclusion reached from these correlation studies, to quote Segal and Barr, is that "formal grammar has no immediate transfer value so far as applied English grammar is concerned."

---

[1] The author is indebted to Mr. Frederick Graham, formerly principal of P. S 210 Brooklyn, Mr Jacob Theobold, principal of P S. 165 Manhattan, Mr Abraham Ehrenfeld, principal of P S 10 Manhattan, and Mrs Ruth D Schatteles, formerly principal of P S 87 Manhattan, for their most helpful cooperation in this experiment, and to the many teachers who participated

[2] Hoyt, F S., "The Place of Grammar in the Elementary Curriculum" *Teachers College Record,* Vol VII, November, 1906, pp. 467–500

[3] Rapeer, L W : "The Problem of Formal Grammar in Elementary Education" *Journal of Educational Psychology,* Vol IV, March, 1913, pp 125–137

[4] Sega, D , and Barr, N R "Relation of Achievement in Formal Grammar to Achievement in Applied Grammar" *Journal of Educational Research,* Vol XIV, December, 1926, pp 401–402

[5] Asker, W · "Does Knowledge of Formal Grammar Function?" *School and Society,* Vol XVII, January 27, 1923, pp 109–111.

However, the true test is the experimental test. What real influence does learning English grammar have on usage? To determine what influence exists, the writer carried out some test-teach-test experiments in Grade VI of four New York City elementary schools

The test which was used in the experiment consisted of forty items and was modeled after the Charters Diagnostic Language Tests Each item (except the first) consisted of a sentence having a grammatical error to be corrected by the pupil by rewriting the sentence in the space beneath the item Samples of the items are.

14 Are you most ready to go?

28. He said that I was most there.

Previous experience with the Charters test indicated that the lowest initial scores, and consequently those items of greater difficulty, were made on adjectives, adverbs, prepositions, and conjunctions. Consequently in planning for this experiment all of the items selected for this test were of these types. This test was called Test I, and was always given first as well as last in the experiment, with the hope of discovering what changes took place on the test due to the activities carried on in between

The technique of the Charters test was retained as the best simple measurable evidence of usage It would have been possible to have made the test a truer measure of choice in an uncontrolled situation, but probably at sacrifice of scorability and reliability On the other hand in using a recognition test, pupils would have been more keenly attentive to differences in form and it would have too narrowly a test situation and not enough an opportunity to recognize spontaneously when a form is incorrect and what the correct usage should be

In every case Test I was immediately repeated at the opening of the experiment, and all gains were computed from the second application of the test instead of the first in order to eliminate the influence of practice on the test The average practice effect was found to be 1 14

Six different experimental procedures were tried. Each procedure was conducted in three classes in a different school under very carefully controlled conditions It was believed that in concentrating a procedure within a school there was less danger of having teachers discuss what they were doing and hence vary their procedure from that which was defined for them No attempt was made to equate groups except the rough one of choosing Grade VI groups This defect in experimental procedure is recognized. The possible influence which differences in selection in the groups might have on the results and attempts to allow for possible differences will be discussed It would have been possible to carry out the study on the scale contemplated if the orthodox equating techniques had been followed.

EXPERIMENTAL PROCEDURE I: REPETITION OF CORRECT FORMS

Four classes in each of two schools, P. S. 10 Manhattan and P S 210, carried out the experiment in pure repetition in 1929. Sheets were prepared which contained forty sentences each of which were grammatically correct. Each sentence contained a usage which which corresponded to a similar incorrect usage in Test I. Examples of items in this sheet of correct sentences (Exercise A) are as follows

14 The boy was almost killed by an automobile.
28 My baby brother is almost two years old

Exercise A was distributed to each child and directions to the teacher stated.

> Have the class read through the sentences aloud and in unison. Let the teacher first read the number of the item and then the class read the sentence. Do not let the children hurry See that they read slowly and distinctly. See that everyone reads Do not allow the class to repeat any of the sentences unnecessarily If the class reads a sentence incorrectly or mispronounces a word, have them repeat it (once) correctly

The sheets were immediately collected when this reading was finished This device made possible one repetition of each correct form by each child

In order to check up on the influences of different amounts of practice on this new test (Test I) as compared with one previous experiment, *one repetition* of Exercise A was given between the two applications of Test I in two classes, *three* repetitions between tests in two classes; *five repetitions* in two classes; and *ten repetitions* in two classes.

This part of the experiment was designed to give a measure of the influence of sheer repetition with which the results of other experimental procedures could be compared

EXPERIMENTAL PROCEDURE II: REPETITION OF CORRECT AND INCORRECT FORMS

In this section of the experiment carried out in P S 165 Manhattan in 1929, mere repetition was again studied, but in this case both the correct and incorrect forms were repeated side by side This new exercise (Exercise A, Form 2) was distributed to the children with directions that it be read aloud and in unison as a class as in the previous procedure. Three repetitions of this exercise were interpolated between the tests and the results were comparable to the three repetitions used in the first experimental procedure

Examples from Exercise A, Form 2:

14. The boy was most killed by an automobile is wrong
14 The boy was almost killed by an automobile is right.
28. My baby brother is most two years old is wrong
28 My baby brother is almost two years old is right.

The pupils read *is wrong* or *is right* after each sentence

EXPERIMENTAL PROCEDURE III· KNOWLEDGE OF DEFINITIONS, RULES, AND PRINCIPLES AND GRAMMAR

In this section, carried out in P. S 87 Manhattan in 1929, an attempt was made to determine the influence on usage (Test I) of mere learning of grammatical rules and principles A grammar study guide was prepared containing definitions, rules, and principles which applied specifically to the errors in Test I.

It was an original intention to include in this first edition of the grammar guide only bare definitions and rules with no illustrations. But this seemed such a formal procedure for children in a classroom even in an experiment and so far removed from school procedure that illustrations were included This experiment does not tell us how much the illustrations in illuminating the rules influenced the results.

Teachers were cautioned as follows:

Do *not* do any of the following
- Do *not* ask pupils to make up examples of their own illustrating the rules.
- Do *not* give the pupils additional examples of the rules.
- Do *not* ask pupils to analyze the rules
- Do *not* give the pupils practice in applying any of the rules

The three teachers cooperating in the experiment were instructed to drill their pupils in memorizing the definitions, rules and principles so far as time permitted. Fifteen or twenty minutes a day for two weeks or more were used

A sample of the grammar guide, as it applies to the particular items already given from the test, is given below

Most and *almost* are often confused. *Most* is an adjective and should modify a noun

Example *Most* boats carry life preservers

Almost is an adverb and should modify a verb

Example· John has *almost* finished his work.

A sheet of "Questions on Rules for Correct Usage of Adjectives, Adverbs, and Conjunctions" was also supplied to each child

Illustrations of items are:

22 What part of speech is *most?*

23. What part of speech is *almost?*

The teachers were instructed to use these sheets with the pupils in drilling them on these definitions, etc

Finally a "Test of Knowledge of Rules for Correct Use of Adjectives, Adverbs, Prepositions, and Conjunctions" was prepared This was to be given to the class after the study of the grammar was finished, but *before* final Test I

Samples of items on this test on grammar rules are

22 Most is an (1) adjective, (2) adverb, (3) preposition, (4) conjunction. ( )

23. Almost is an (1) adjective, (2) adverb, (3) preposition, (4) conjunction ( )

The teachers were given the express goal of preparing their pupils for this test on grammar rules, using every means at their command, direct or indirect They were to aim at one hundred per cent on the grammar rules test if possible and the extent to which they reached this goal was a measure of their success in teaching the grammar rules. That they were remarkably successful in this is evidenced by the fact that out of a total possible score of 36 on this test, the three classes *averaged* 33 1, 24 2, and 30.1.

The purpose of this work was to discover what influence the learning of the rules had on Test I

EXPERIMENTAL PROCEDURE IV· GRAMMAR ANALYSIS

The fourth experiment was carried out in one class in P. S 87 Manhattan and two classes in P S 210 Brooklyn in 1930 In this part of the experiment the value of ability in the grammatical analysis of sentences in improving English usage was studied. Experiment IV went one step further than Experiment III and gave pupils practice in analyzing the grammatical construction of certain words and phrases A new grammar guide was made which included not only the definitions, principles and rules as before, but also certain exercises in which pupils were expected to *recognize* and *name* the construction In Experiment IV teachers were instructed *not* to do any of the following:

> Do *not* have the pupils memorize or learn the rules or principles
> Do *not* give examples of or emphasize the *wrong* form in any of the exercises
> Do *not* give the pupils practice in applying any of the rules
> Do *not* have the children make up illustrations of their own.
> Everything should concentrate on having children learn to analyze the sentences given

Besides the definitions and examples in the grammar manual which are similar to those used in the previous experiment exercises were included of which the following are typical:

1 *Most* boats carry life preservers *Most* is an______________
and modifies ______________
2 John has *almost* finished his work *Almost* is an______________
and modifies______________.

A test of thirty-four items with a total possible score of sixty-one on "Grammar Analysis" was prepared and the teachers cooperating in this experiment were instructed to do everything possible to prepare their pupils to do well on this test. Their success is evidenced by average scores of 37.9, 42 6, and 42 8.

EXPERIMENTAL PROCEDURE V. CHOICE OF CORRECT CONSTRUCTIONS

Experiment V, which was conducted in P S. 10 Manhattan in 1930, tested the value of practice in choosing correct constructions on English usage as measured by Test I

A new grammar guide was again prepared including not only the necessary definitions, principles, rules, and illustrations, but also exercises in which pupils were given practice in choosing correct constructions Samples of these exercises on the same distinction *most* and *almost* are·

Exercises on Use of Most and Almost—Insert *most* or *almost* in the following blank spaces:

1 Are you____________ over your cold?
2. ____________birds build nests
3 Which is the____________ fun?

In this experiment the teachers were instructed to limit their teaching strictly to giving the pupils practice in choosing the correct constructions. They were told *not* to have the pupils memorize or learn rules or principles, *not* to give examples or emphasize the wrong form in any of the exercises, *not* to analyze the constructions in any of the examples or exercises, and *not* to have the children make up illustrations of their own

As before, a test was prepared on the choice of correct constructions and teachers were expected to help prepare their pupils directly for this test. Items in the test were of the same kind as the practice exercises, illustrations of which have already been given. The success of the teaching may be seen from average scores of three classes The scores were 24.3, 24 4, and 26 7 out of a possible 33.

In this procedure, as in the other, Test I was given first and last, the aim being to see what effect the practice in choosing correct constructions has on Test I.

EXPERIMENTAL PROCEDURE VI: WHOLE PROGRAM

In 1930 the whole program comprised in the previous experiments was tried out in P S. 165 Manhattan After Test I, the grammar manual was distributed and pupils in three Grade VI classes learned the various definitions, principles, and rules, practiced analyzing sentences for the constructions of certain words, and practiced choosing correct constructions. When all this was done tests were taken on knowledge of grammar rules, ability to analyze certain constructions, and ability to choose correct forms This was capped by three repetitions of the right-wrong forms of Exercise A, Form 2, described in Procedure II Then Test I was repeated. The average scores on the tests for three classes were as follows

TABLE I.—AVERAGE SCORES ON KNOWLEDGE, ANALYSIS AND CHOICE TESTS GIVEN TO THESE CLASSES IN PROCEDURE VI

	Knowledge test	Analysis test	Choice test
Total possible score	36	61	33
Class VIA	34 2	55 4	31 4
Class VIB$_1$	30 0	45 8	27 6
Class VIB$_3$	30 1	33 1	26 7

TABLE II.—SCORES MADE ON TEST I BEFORE AND AFTER AN EXPERIMENTAL PROCEDURE TOGETHER WITH GAINS BY CLASSES AND AVERAGE GAINS

Experimental Procedure I practice	One repetition		Three repetitions		Five repetitions		Ten repetitions	
	P S 10 VIA$_3$	P. S 210 VIB$_{4R}$	P S 10 VIA$_1$	P S 210 VIA$_1$	P S 10 VIB$_4$	P S 210 VIB$_1$	P S. 10 VIB$_2$	P S 210 VIA$_3$
Second test	5 80	15 87	10 44	11 69	8 33	22.45	6 48	7 59
First test	5 40	15 55	9 22	12 24	7 04	20 75	5 70	7 71
Gain	+ 40	+ 32	+1 22	− 55	+1 29	+1 70	+ 78	− 12
Average	+ 36		+ 34		+1 50		+ 33	

Experimental Procedure II, practice on right and wrong, P. S 165	VIB$_1$	VIB$_2$	VIB$_3$
Second test	24 77	21.70	24 48
First test	15 00	12 26	13 36
Gain	+ 9 77	+ 9 44	+11 12
Average gain 10 11			

Experimental Procedure III, learning rules, P S. 187	VIB$_1$	VIB$_3$	VIIA$_1$
Second test	25 44	13 37	23 80
First test	17 20	10 26	17 48
Gain	8 24	3 11	6.32
$\sigma_{\text{mean gain}}$	82	73	75
Average gain 5.89			

TABLE II —*Continued*

Experimental Procedure IV, analysis, P S 87 and P S 210		P S 87 VIB	P S 210 VIB_{4R}	P S. 210 VIB_{5R}
Second test		23 90	16 28	15 02
First test		19 83	11 00	12 23
Gain		4 07	5 28	2 79
$\sigma_{\text{mean gain}}$		84	59	59
Average gain	4 05			

Experimental Procedure V, choice, P S 10		VIB_1	VIB_3	VIB_4
Second test		24 15	17 71	18 40
First test		14 50	10 16	11 58
Gain		9 65	7 55	6 82
$\sigma_{\text{mean gain}}$		55	63	98
Average gain	8 01			

Experimental Procedure VI, total program, P S 165		VIA_1	VIB_1	VIB_2
Second test		29 40	28 72	26 20
First test		13 45	16 03	14 72
Gain		15 95	12 69	11 48
$\sigma_{\text{mean gain}}$		76	71	86
Average gain	13 37			

Before discussing the significance of the results consideration must be given to possible extraneous factors which may have influenced the results The groups were not equated, as was previously mentioned, except by the rough method of selecting Grade VI classes (one class was a VIIA section) It may be seen by inspecting the average scores on the first test in the above table that there are wide differences in initial language usage It is possible that there are also differences in general ability, and it is concensible that the brighter classes might gain so much more from the experimental procedure than duller classes by virtue of their brightness as to invalidate the apparent differences.

These possibilities were investigated by finding the correlations which exist between various factors.

TABLE III.—COEFFICIENTS OF CORRELATION BETWEEN MENTAL AGE, IQ, INITIAL SCORE ON TEST I AND GAIN ON TEST I

	Initial score-gain	MA-initial score	IQ-initial score	MA-gain	IQ-gain
Experimental Procedure III (*rules*) (average three classes)			+ 072		+ 084
Experimental Procedure IV (*analysis*)					
(Average three classes)			+ 330		+ 157
(Average two classes)		+ 361		+ 076	
Experimental Procedure V (*choice*) (one class)			+ 415		+ 152
Experimental Procedure III (*rules*)					
(Average three classes)	— 135				
(*Analysis*) IV (average three classes)	—.271				
(*Choice*) V (average three classes)	— 440				
(*Total*) VI (average three classes)	— 334				
Average	— 295				
Correlation for total group	—.138				

These correlations indicate that there is a marked negative correlation between initial score on Test I and gain on the test In other words pupils, making the lowest initial scores on the test make the greatest gain. This is probably an artifact of the test. Pupils with the higher scores do not have room enough on the test to make as great gains as those making lower scores. That this factor had no appreciable effect on the average gains *by classes* is seen from the following table If this factor did influence average gains by classes, then the

TABLE IV.—AVERAGE INITIAL SCORE AND AVERAGE GAIN FOR EACH OF THE EXPERIMENTAL PROCEDURES

School	Experimental procedures		Average initial score	Average gain
P. S 10.. P S. 210.	Practice	One repetition	10 48	+ 36
		Three repetitions	10 73	+ 34
		Five repetitions	13 89	+ 1 50
		Ten repetitions	6 71	+ 33
P S. 165	Practice: Right and wrong		13 54	+10.11
P S. 87	Rules		14 65	+ 5 89
P. S 87 and P. S 210	Analysis		14 35	+ 4 05
P S 10	Choice		12.08	+ 8 01
P S 165	Total program		14 73	+13 37

differences are really greater than they appear, for in this table there is a tendency for those classes with the highest initial average score to make the most gain.

There is also a marked correlation between intelligence and initial score in Test I. The brighter pupils tend to exhibit better language usage as measured by Test I. The correlation of gain and intelligence is negligible although it tends to be positive. Were it not for the artifact of the negative correlation between initial score and gain it is probable that the correlation between intelligence and gain would be much larger. Semi-partial correlations were computed between gain and IQ, eliminating initial score from gain,[1] showing that the correlation between gain and IQ is lower than it would otherwise be due to the nature of the test.

TABLE V.—SEMI-PARTIAL CORRELATIONS BETWEEN GAIN AND IQ ELIMINATING INITIAL SCORE FROM GAIN

School	Class	Correlation IQ-gain	Correlation IQ-gain with initial score eliminated from gain
P. S. 87	VIB_1	−.337	+.051
	VIB_3	+.676	+.533
	$VIIA_1$	−.080	+.158
P. S. 87	VIB	+.379	+.419
P. S. 210	VIB_{4R}	−.046	+.235
	VIB_{6R}	+.137	+.163

The influence of general ability of a group may have had some influence on the gains, but probably not much. The correlation is positive but low. The classes in P. S. 165 had high IQ's on the average and in both cases made the highest gains. On the other hand both

---

[1] By a semi-partial correlation is meant the elimination of the effect of a third variable from one of the variables. In this case the formula used was

$$r_{(12)3} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{1 - r_{12}}}$$

where

1 = gain
2 = Initial score
3 = IQ

Semi-partial coefficients of correlation were first developed by Franzen. See Franzen, R. A Comment on Partial Correlation. *Journal of Educational Psychology*, Vol. XIX, 1928, pp. 194–197.

of the classes in P. S. 210 participating in Experimental Procedure IV (analysis) were rapid advancement classes with high IQ's but they made comparatively low gains Again the classes in P. S. 10 participating in Experimental Procedure V (choice) had average IQ's which were really low in this experiment and yet their gains were high. It is not believed that the fact that classes differed in initial ability had appreciable influence on the results.

INTERPRETATION OF THE RESULTS

1. The results show that mere repetition of correct forms cause small gains on the test It is possible that the small gains of 0 3 for one, three and ten repetitions of the correct forms could be explained as mere practice effect The original practice effect between the first and second trials on Test I was 1.14, and it is possible that there should be another small practice effect gain between the second and third trials At all events, when compared with the other procedures mere repetition of correct forms is very inferior.

2 If one recalls the gains found for *motivation* as described in a previous experiment[1] it will be seen that they sink into greater insignificance than appeared in the earlier article. Motivating devices apart from the learning experiences offered are practically impotent

3. It will be seen from Experimental Procedure IV that mere learning of definitions, rules and principles of grammar yielded an appreciable gain on the usage test This result was wholly unexpected and is hard to explain. It was thought before beginning the experiment that mere memorization of grammar rules had little or no influence on usage The experiment, however, indicates that somehow or other these Grade VI children extracted enough meaning from the formal and abstract definitions and rules with which they were previously unacquainted to make more correct changes on the test than they did the first time. The positive correlations of gains with intelligence would indicate that this "transfer" was one that the brighter children profited by more than the duller children Within the limited use of language as found in this experiment it is possible to say that mere learning grammar rules *does* improve usage

4 Training in the analysis of grammatical constructions and practice in the choice of correct constructions each taken alone caused improvement in English usage The development of ability to analyze

---

[1] Symonds, P M and Chase, Doris Harter. Practice vs Motivation *Journal of Educational Psychology*, Vol. XX, 1929, pp 19–35.

sentences for grammatical constructions apparently had about the same influence on usage that memorization of the rules had. Practice in choice of correct expressions had measurably greater influence than the learning of rules or analysis

5 Mere repetition of the right and wrong forms in succession so that the distinction between the two is clearly brought out, and so that it is definitely stated which is the right form, had greater influence on usage than any of the work with grammar.

6 The whole program including memorization of rules, practice in analysis of grammatical constructions, choice of correct forms and mere repetition of correct and incorrect forms in succession yielded results in improved usage better than any one single method alone It is estimated that these Grade VI classes made a gain in correct usage which is ordinarily to be expected in three years by working on this program fifteen minutes a day for three weeks. The experiment has not included subsequent testing to determine the permanence of these gains Presumably there is the usual falling away and the re-adoption of habits prevalent in home and community

DISCUSSION

One thing stands out clearly. It is the quality of drill or practice that counts and not its amount. Mere mechanical repetition apparently yields almost no learning. Instead of increasing the amount of drill in school subjects, more attention should be paid to the nature of the drill The best results come by making what is to be learned easily identifiable as Thorndike has emphasized in recent lectures

This study should not necessarily be interpreted as a justification of the teaching of English grammar Grammar does have an influence on usage, but at what cost? Without doubt for most children the difficulty and trouble of learning grammar as a means for improving usage is so great that more direct attacks on usage are certainly more profitable Our own results show that mere repetition of correct forms, where it is clearly indicated what is correct and what the critical point at issue is, has more influence on usage than any procedure with grammar.

For gifted children the study of grammar may be profitable In the first place, the gifted child learns grammar relatively more easily. In the second place, with general principles to work from, a knowledge of grammar will cover more cases Whoever really comprehends a generalization has command over a much wider range of specific applications. But the average child finds it difficult to learn the definitions, rules, and principles, and finds it difficult to apply them when learned. For most children learning correct usage is more profitably accomplished when the specific items are learned directly The child who has learned by direct practice to distinguish right from wrong usage is probably better equipped to learn the grammar which gives a logical explanation of the habits of correct usage which he has formed. In this sense grammar is a summary or epitome of the usage which has already been learned directly, rather than a tool for guiding the learning of correct usage. In the writer's opinion it is relatively unprofitable for the average child either to study grammar as a means for learning correct usage (because there are more direct, simpler, and easier means for learning usage) or as a means of summarizing the correct usage which one has learned (because there are other things more worthwhile in the curriculum).

In all the experimental work it is tacitly understood that pupils are making some effort to do well in school, to do well on tests, to please the teacher, to do what other children are doing, to learn new things, and the like Unless there was some driving force the pupils would never have taken the tests, or attended to or learned the exercises Our previous experiment showed that additional stimuli to learning cannot be imposed from without. But there must be a minimum of driving force which can be depended upon to direct the children's attention to what is to be learned. The selection of material in harmony with pupil's *interests* and *previous learnings* is certain to produce better learning results than when interests and previous learnings are disregarded.

It is always questionable as to how far the results of a single experiment can be extended to apply to dissimilar situations. The writer likes to believe that these experimental results apply to problems of conduct in general This experiment demonstrates that mere learning rules or codes has a direct influence on conduct, more for the brighter children than for the duller children. Apparently these children extract enough from the formal rules to apply to their own behavior The practical problem, however, is not Can rules function in conduct? but What is the most economical method of guiding or controlling conduct? This depends partly on the person's ability to comprehend generalizations and apply them to specific situations. Probably for the majority of people the simple direct method of telling them what is the "right" or best thing to do will produce the best results. And if it is not known or not certain that one thing rather than another is right or best, or if it depends on the circumstances, then it is preferable either in the first instance to teach nothing, or in the second instance to give considerable help in solving the problem which will fit the response to the circumstances.

Finally, a word as to the experimental attack on problems of this type as compared with correlation methods Correlation analysis is essentially static. It merely reports the results of whatever methods or combination of methods have produced the present situation. Correlation analysis does not begin to exhaust the possibilities in method—it does not even experiment with method. Correlation analysis is extremely limited insofar as its guidance for education is concerned The point has been reached where the correlation method has about exhausted its possibilities. The time has come for a more earnest use of the experimental approach to problems of learning and education.

However, if the experimental method is to be used it must give heed to the law of the single variable Every true experiment in education presents an artificial educational situation. Time and again during the course of the work reported in this article, principals and teachers would complain that the methods which were being used were too formal and dry to be practically used. To this I would agree To most persons an educational experiment must make use of the best conceived organization of methods and materials possible. Not so. The actual educational situation is too complex to be used in experiment. When one has compared one actual classroom method with another, one does not know to what specific factor to attribute the results In true experimental work a single factor must be pulled out of its setting and alone be allowed to vary Only in this way is it possible to build up educative procedures that are based on correct principles

A SCALE OF MILITARISM-PACIFISM[1]

D D DROBA

University of Chicago

THE PROBLEM

The purpose of this article is to describe the construction of a scale for the measurement of militarism-pacifism The term "militarism-pacifism" is used here to specify a particular attitude. In a very broad sense it denotes a predisposition to act with reference to the issue of war *vs.* peace.

To measure militarism-pacifism directly is impossible One cannot enter into an immediate contact with the predispositions of other persons. As an indirect measure of attitude the verbal expression of a person, in the form of statements, was chosen The statements cover the following topics in this field: Causes of war, purposes of war, results of war and peace, what is to be done at present about war and peace, what is to be done in case of war, and general judgments about peace and war.

We can assume that there is some relation between action and verbal expression. That is, a person who endorsed militaristic statements would probably tend to act like a militarist, and a person who endorsed pacifistic statements would be likely to act like a pacifist. However, no assumption is made as to the extent of relation between the verbal expression and action. If it is discovered that a person has endorsed statements indicating a pacifistic attitude, one cannot be sure that he will also act exactly in accordance with his endorsements

In the present study verbal expressions or opinions will be used to designate the successive steps on the militarism-pacifism scale In constructing the scale it will be desirable to find some way of determining the distances between the successive steps on the scale. It will be desirable, furthermore, that the distances between the steps be approximately equal. How to equalize these distances constitutes the chief problem

To summarize, the main problem of this experiment is to construct a scale of militarism-pacifism with approximately equal steps on the scale—the steps to be represented by statements of opinion concerning war and peace.

---

[1] The scale, in a modified form, together with instructions for its use, was published by the University of Chicago Press.

THE METHOD

A set of two hundred thirty-seven statements expressing various degrees of militarism-pacifism was collected from several sources. These sources may be divided into three main divisions. The literature on war and peace, statements about the issue of war *vs* peace written by one hundred undergraduates and twenty graduates, and a number of statements devised and modified by the writer Out of the two hundred thirty-seven statements sixty-seven longest, least clear, and least relevant statements were discarded Three professors and the writer then went over the remaining statements, and as a result forty more statements were eliminated, leaving one hundred thirty statements for use in the experiment.

Three hundred students attending the University of Chicago were used as subjects. Among them were eleven unclassified, eleven freshmen, ninety sophomores, fifty-seven juniors, thirty-nine seniors, forty-two year graduate students, thirty second year graduates, seventeen third year graduates, and three Ph.D.'s. Most of the subjects were students taking elementary psychology Other departments which also assisted were sociology, political science, divinity school, and education Only voluntary service was asked for except in a few cases when the instructor made the experiment a substitute for some class exercise.

The presentation of the material was made in some cases by the experimenter, in others by an instructor. Generally the envelopes containing the material were distributed first during the class hour The experimenter then read the instructions while the subjects followed him in reading their own sheets. Any questions asked by the students at this time were answered. The subjects were then asked to do the task at home and to return the envelopes with the material complete in a few days

When the student opened the envelope he found in it the following material: A sheet of instructions, one hundred thirty statements each on a separate slip 5½ inches by 4½ inches, eleven index slips of the same size, each labeled with a Roman number I–XI, and a slip of similar size for obtaining the necessary information about the subjects.

The subject was instructed to lay out before him the eleven slips. He then was asked to put on slip I those statements which he believed express the most extremely militaristic opinions On slip XI he was requested to put those statements which he believed express the most extremely pacifistic opinions, and on slip VI to put those which he thought express neutral opinions On the rest of the slips he was instructed to arrange the statements in accordance with the degree of militarism and pacifism expressed in them.

Only about seventy-five per cent of the envelopes were returned This automatically eliminated some of the most careless individuals A few envelopes were returned unused which was preferable to a thoughtless and hurried performance. The index slips and the instruction sheets were returned in most cases and were used over again with the newly prepared sets of statements.

The frequencies of judgments obtained from the three hundred students for each statement were tabulated. Table I shows the frequencies of judgment only for statements chosen to constitute the final scale. These frequencies were then cumulatively added and the percentages of the total three hundred judgments determined for each of the resulting sums Three samples are shown in Table II It is evident from the second column that statement number 9 was put in group I (first column) by two hundred fifty-four subjects. Two hundred fifty-four is 84.6 per cent of the three hundred subjects. Twenty-eight people put the statement in group II Adding twenty-eight to two hundred fifty-four, we get two hundred eighty-two which is ninety-four per cent of three hundred. Seven people put the statement in group III. Seven and two hundred eighty-two is two hundred eighty-nine; that is, 96.3 per cent of the total number of subjects or judgments. Percentages thus obtained were called cumulative proportions

The same method of calculation was used for all the one hundred thirty statements In order to show how this method applies to all kinds of distributions the three statements, 9, 10, and 8, were purposely selected to represent three completely different distributions. Statement 9 was put by most people to the extreme left, statement 10 in the middle groups, and statement 8 in groups lying to the extreme right (see Table II)

From data such as found in Table II the cumulative proportions were plotted against the eleven groups or degrees of militarism-pacifism. Sample curves are shown in Fig. 1 Militarism-pacifism

TABLE I

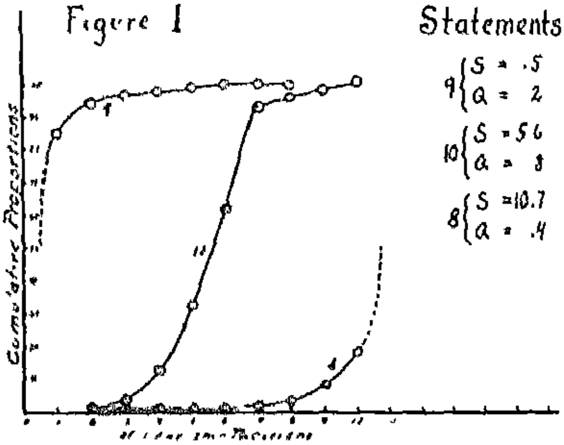
Statements	Groups										
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
1	11	60	81	80	51	10	1		1	2	1
2			2	1	1	2	15	32	50	02	00
3	.	3	25	31	115	07	22	5	1	1	
4		2	2	5	2	8	16	80	85	41	20
5	1	4	1	3	1	10	18	30	70	02	50
6		1	14	40	106	27	50	31	14	10	2
7	115	03	35	20	5	. .		1		1	
8		1	.		.	. .	2	5	15	31	210
9	251	28	7	3	4	3					
10	.	2	8	20	60	88	93	0	0	8	
11	2	3	7	13	28	20	180	33	9	5	
12	13	36	17	86	80	33	1			3	1
13	6	18	38	09	105	54	4	2	1	3	
14		2	1	5	16	123	55	34	32	20	0
15	5	1		.	1	1	4	8	13	55	212
16	2	1	3	0	10	45	78	70	40	24	15
17	102	87	58	30	13		2		2		
18	1	1				2	14	27	38	70	147
19	73	74	77	10	26	3		1			
20	31	60	79	05	42	8	1	1	.		1
21		1	2	2		10	38	65	90	72	24
22		.		. .	3	22	78	71	72	40	11
23	.		2	1	1	1	22	40	81	84	50
24		7	21	60	131	55	11	2	1	3	
25	. .	.	.	1	3	10	05	70	55	43	11
26				.		1	4	10	10	58	208
27	42	05	78	02	31	13		.	2	3	1
28	104	67	21	13	4	1					
29				1	2	7	43	00	75	02	44
30	65	70	00	52	20	7	.	1	2	1	1
31	33	50	72	08	59	13	1	2		2	
32	.	2	1		5	20	53	73	74	40	14
33			1	1		2	23	25	58	85	105
34	2	1	2	5	10	103	35	32	20	18	6
35							4	12	18	40	220
36	1	17	44	70	100	35	10	2	2	4	
37		2	2	11	30	65	08	50	23	15	1
38	. .		1		2	4	21	18	33	65	150
39	03	85	62	30	20	1			1		2
40		.		2	4	20	42	67	78	62	16
41	10	40	04	82	70	14	3	2	1	2	
42	118	81	43	13	12		.		2	1	
43		1	3	1	3	07	80	51	46	30	0
44	0	8	19	22	35	208	2				

TABLE II.—CALCULATION OF CUMULATIVE PROPORTIONS

Number of the statement	9		10		8	
Groups	Cumulative frequency	Per cent	Cumulative frequency	Per cent	Cumulative frequency	Per cent
I	254	84.6				
II	28		2		1	
	282	94		.6		.3
III	7		8		0	
	289	96.3	10	3.3	1	.3
IV	3		26		0	
	292	97.3	36	12	1	.3
V	4		60		0	
	296	98.6	96	32	1	.3
VI	3		88		0	
	299	99.6	184	61.3	1	.3
VII	0		93		2	
	299	99.6	277	92.3	3	1
VIII	0		9		5	
	299	99.6	286	95.3	8	2.6
IX	1		6		15	
	300	100	292	97.3	23	7.6
X			8		31	
			300	100	54	18
XI					246	
					300	100

was represented on the base line by the eleven groups running from extremely militaristic groups to extremely pacifistic ones The cumulative proportions were located on the y-axis running from zero to one hundred.

Sample curves are shown in Fig 1 Militarism-pacifism was represented on the base line by the eleven groups running from extremely militaristic groups to extremely pacifistic ones The cumulative proportions were located on the y-axis running from zero to one hundred



From the curves thus obtained two values were determined: The scale value and the quartile deviation. The scale value of a statement was found as follows Somewhere on the chart the curve intersects the fifty per cent horizontal line, that is a line running horizontally at the level of fifty per cent and at right angles to the y-axis. From the point of intersection a line was dropped at right angles to the x-axis The point where this line intersects the base line determined the scale value of the statement The scale value of statement 10 (see Fig. 1) was found to be 5.6 and recorded on the right-hand side of the figure A similar procedure was followed in finding the scale value of each of the one hundred thirty statements

With some of the extreme statements, however, the curve did not reach the fifty per cent line Examples of this may be found in Fig. 1 and Fig 3. In such cases the curve was projected at sight down

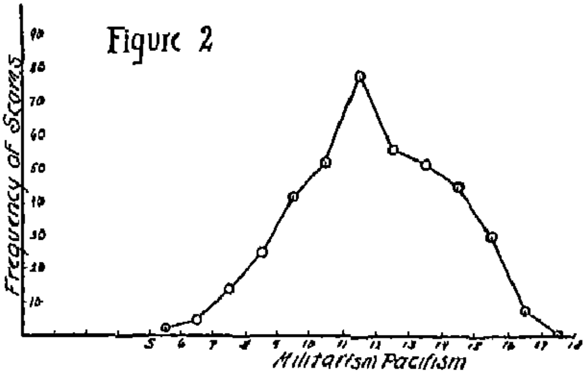


Figure 2

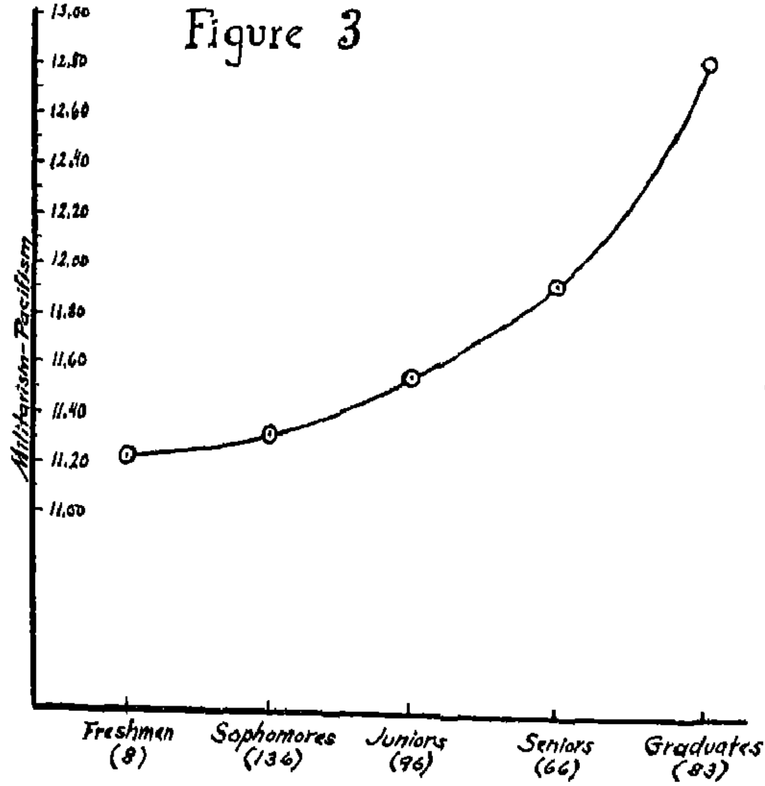


Figure 3

or up as far as the fifty per cent line. Naturally, such procedure is not quite as accurate as when the curve reaches the fifty per cent line, but it is the best method available for the extreme curves.

Variability of a statement was measured by the semi-interquartile range or quartile deviation, designated by *Q*. The greater the value of *Q*, the more variable the statement, because people disagree more as to where the statement belongs. *Q* value was obtained by simply reading off the range on the base line corresponding to the distance between the two points where the curve intersected the twenty-five per cent and the seventy-five per cent horizontal lines, and dividing this distance by two. The variability of statement 9 was found to be 0.2, the variability of statement 10 was 0.8, and that of statement 8 was 0.4 (see Fig. 1).

THE SCALE

1. *Construction.*—The next task was the selection of statements to constitute the final scale. Two major criteria used for this purpose were (*a*) the scale values and (*b*) the *Q* values. All of the one hundred thirty statements were arranged first on the basis of the scale values in ascending order, ranging between 0.5 and 10.7. Nearly every step in this range was represented by several statements of practically identical scale values. All statements of identical scale values were arranged according to their *Q* values.

It was decided that there would be twenty-one steps in the entire scale, each step to be represented by one statement. An additional statement was to represent the origin. Approximately two statements were picked from each of the eleven groups. But in so doing the experimenter was guided by the scale values of each statement rather than by the groups.

In case of identical scale values the least variable statements were picked as measured by the *Q* value. In a few instances both the scale values and the *Q*'s were identical. Two minor criteria were used in cases of this type. The brevity of a statement and the shape of the curve of its distribution. A regular curve was preferred to a less regular one and a short statement to a longer one.

As a result, a series of twenty-two statements were obtained that were about equally spaced along the scale. The distance between two adjacent statements seemed to the three hundred judges, on the whole, as large as the distance between any other two adjoining statements. In terms of scale values none of these distances was larger than .6, and none smaller than .4 with the exception of the distance between the last two statements, which was .3. A second form of the scale was selected by examining the scale values and the Q's again and picking equally spaced and least variable statements. Twenty-two statements were selected to match the first form as nearly as possible.

The scale was readjusted on the basis of the frequencies of endorsement obtained from four hundred students. The frequencies of endorsement were cumulatively added, and each sum in one of the forms was compared with the corresponding sum in the other form. If one of the sums was larger, the statements were exchanged so as to balance the frequencies in one form with the frequencies in the other

After the two forms were thus equalized the scores were calculated for each and all of the four hundred subjects The distributions of the frequencies of scores obtained from the two forms are shown in Fig 2

2. *Description.*—Forty-four selected statements constituted the main body of the scale in two forms. Thus each of twenty-one steps on the scale was represented by a pair of statements One pair, the most extremely militaristic statements, represented the origin and were designated by *O*. The two most extremely pacifistic statements were designated by number twenty-one. In order to assure greater attention and careful reading on the part of the subjects, the statements were arranged in random order. Written instructions were found at the beginning of the scale. Attached to the scale was the information sought about those to whom the scale was given

The instructions, the scale statements, and the information section are reproduced below. In front of each statement code numbers are found. After each statement in the parenthesis three numbers are given: The scale value, the quartile deviation, and the equivalent number (about double the scale value) of that statement. The latter number was used in scoring. In giving the scale these three numbers are, of course, omitted.

A Scale of Militarism-Pacifism

Instructions

The purpose of this experiment is to measure differences in opinions relating to peace and war Below you will find a number of statements expressing various degrees of opinion about militarism-pacifism.

If, on the whole, you agree with the sentiment of a statement, mark it with a *plus sign* in the parenthesis.

If, on the whole, you disagree with the sentiment of a statement, mark it with a *minus sign* in the parenthesis

Please vote for each statement even if it seems ambiguous and you have to make some sort of guess

Only mass results will be published which will not distinguish individual opinion.

Statements

( ) 1. Multitudes are benefited by learning the lesson of war-time discipline. (3.0, 0.0, 5)
( ) 2. War in the modern world is as needless as it is suicidal. (9.5, 0.9, 18)
( ) 3. Those military units should be retained which afford training to the body and mind. (4.8, 0.6, 8)
( ) 4 Army discipline injures self-respect and individuality. (8.0, 0.9, 15)
( ) 5 The losses in human life and property, great as they are, are small evils compared to the undermining of morals and the lowering of standards of culture and civilization caused by war (9.0, 0.9, 17)
( ) 6. Wars are justifiable only when waged in defense of weaker nations (5.2, 1.1, 9)
( ) 7 War is ennobling and stimulative of the highest and best qualities of humankind (1.0, 0.8, 1)
( ) 8. It is the moral duty of the individual to refuse to participate in any way in any war, no matter what the cause. (10.7, 0.4, 21)
( ) 9. Might is right (0.5, 0.2, 0)
( ) 10. The most that we can hope to accomplish is the partial elimination of war. (5.0, 0.8, 10)
( ) 11 The evils of war are slightly greater than its benefits. (6.5, 0.5, 12)
( ) 12. Until an equivalent discipline is organized war must have its own way. (3.0, 1.0, 6)
( ) 13 Because right may be more important than peace, war may be the lesser of two evils (4.2, 0.9, 7)
( ) 14 Along with patriotism world-citizenship should be taught in all secondary schools. (6.0, 1.0, 11)
( ) 15. Persons should pledge themselves never to aid any future war (10.4, 0.5, 20)
( ) 16 It is almost impossible to have a large military force without being tempted to use it (7.0, 1.0, 13)
( ) 17. War is the tonic of races (1.5, 1.1, 2)
( ) 18 When the next war breaks out we should tell the diplomats who lead us to it that we will not follow them (10.0, 1.1, 19)
( ) 19. Militarism is necessary for the proper defense and protection of the individuals of a country (2.0, 1.0, 3)
( ) 20. So long as any people, white, black, brown or yellow, hold weapons in their hands, we must not commit the folly of disarming. (2.5, 1.0, 4)
( ) 21. Under the scourge of war a nation has no opportunity for cultural development (8.4, 0.9, 16)

( ) 22 The soldier suffers tremendously and gains very little (7 0, 1 0, 14)

( ) 23 The evils that war brings in its train far outweigh any possible benefits (8.9, 0 9, 17)

( ) 24 We should have a moderate amount of military training in our schools, (4 4, 0.6, 8)

( ) 25. No scheme of aggression or conquest can be pursued for any considerable length of time without enfeebling victor as well as vanquished. (7 4, 1 1, 13)

( ) 26 When war threatens we should refuse the call to service and increase our anti-war activity (10 5, 0.7, 20)

( ) 27. It is foolish to talk of the abolition of war, since pugnacity is an ineradicable human instinct (2 5, 1.0, 4)

( ) 28 There is no progress without war (0 6, 0.7, 0)

( ) 29. Militarism should be abolished from the curriculum of the state schools (8.4, 1 1, 16)

( ) 30 It is not in war but in peace and prosperity that our worst vices develop and grow rank (2 1, 1 1, 3)

( ) 31. We cannot hope to do away with war, because it is part of the unending struggle for survival in a crowded world (2.9, 1 1, 5)

( ) 32 If armed conflict between individuals and cities can be outlawed, it is possible to outlaw armed conflict between nations (7 8, 1 0, 14)

( ) 33 Every war shows cowardice, murder, arson, graft, and leaves a trail of personal and national demoralization. (9 5, 1 0, 18)

( ) 34 The most frequent cause of war is the rivalry of nations for possession of territory, markets, concessions, and spheres of influence (5 6, 0 9, 10)

( ) 35 There is no conceivable justification for war (10 7, 0 7, 21)

( ) 36 Military training is imperative, but it should be voluntary. (4 1, 0 8, 7)

( ) 37 Nations should agree not to intervene with military force in purely commercial or financial disputes (6 3, 0 9, 11)

( ) 38 Concerning war we must be abolitionists. (10 0, 1.1, 19)

( ) 39. The abolition of war would mean effeminacy, softness, debilitation, and degeneracy (1.6, 1 0, 2)

( ) 40. A host of young men entered the war in a spirit of idealism and unselfish devotion to a great cause, only to return disillusioned and cynical as to the value of ideals (8.1, 1.0, 15)

( ) 41. For the liberty of oppressed nations wars should be fought (3 4, 1 0, 6)

( ) 42. Compulsory military training should be established in all countries (1 0, 0.9, 1)

( ) 43 Pugnacity, rivalry and self-interest are natural, but need not result in war any more than human desire for dominance need result in slavery (6 8, 1 1, 12)

( ) 44 Peace and war are both essential to progress (5 4, 0 6, 9)

Information

Name
Classification (underline one)
Freshman Junior Graduate
Sophomore Senior
Sex (underline one). Male Female
Age

3. *Scoring.* For scoring either the original scale values or the equivalent, numbers might be used. In this experiment for the purpose of simplification and because the original scale values were fairly evenly distributed, equivalent numbers were used as a basis for scoring. The score was the average of the equivalent numbers of the statements marked plus.

4. *Reliability.* For calculation of the reliability of the scale the method of form comparison was chosen. For this purpose the data obtained from four hundred students (see application of scale) were used. To correlate the two forms the product moment coefficient of correlation was used. The correlation was found to be +.83. The Spearman-Brown formula when applied to this gave a value of +.90, which is the expected correlation of the two forms if correlated with two other forms. The latter correlation is to be regarded as the reliability of the total scale as used in its application.

APPLICATION OF THE SCALE

The constructed scale was given to four hundred students attending the University of Chicago. The four hundred students included eight freshmen, one hundred thirty-six sophomores, ninety-six juniors, sixty-six seniors, eighty-three graduates, and eleven unclassified students. The ages of these subjects ranged between seventeen and forty-four, the average being 21.8. Scores on the scale were studied in relation to four factors: Education of students, scholarship, sex, and church affiliation.

1. *Education of Students.* The four hundred records were arranged on the basis of the educational status of each student. Records of freshmen were put in one group, records of sophomores in another, and so on. All in all, five groups were obtained. The types of groups and the number of students in each are shown in Table III. For each of the groups an average of the scores was then calculated, also the

standard error of the average, and the standard deviation of the distribution of the scores.

A tendency may be observed in the successive differences between the averages of the five groups. If taken separately they are hardly of any significance, but when compared one can observe a certain definite continuity from one group to another. These differences all run in the same direction, which indicates a tendency toward greater pacifism as education increases. This increase is graphically represented in Fig. 3.

TABLE III

No	Class	*N*	Mean score	σ_m	σ
I	Freshmen	8	11 23	71	2 02
II	Sophomores	130	11 32	20	2 35
III	Juniors	96	11 50	20	2 05
IV	Seniors	66	11 03	28	2 30
V	Graduates	83	12 82	23	2 10

The second feature of the differences is their gradual increase. This function, somewhat magnified, may also be seen in Fig. 3. A curve drawn through the averages appears to be postively accelerated, the scores ranging from the mean score of the freshmen class to the mean score of the graduates.

These results tend to indicate that education has something to do with the formation and modification of attitudes toward war and peace. With educational advancement students tend to become more favorable to a peaceful way of solving international differences. This general trend is seen in Table III and Fig 3, according to which there is a continuous change in attitude toward pacifism as the students advance in the college to the graduate school.

2 *Scholarship*.—Scholarship standings were obtained from the Bureau of Records for two hundred twenty-seven students. The undergraduate grades were expressed by the number of points. But the graduate grades had to be converted into numerical values by the experimenter This was done as follows: $H = 6$, $P = 4$, $U = 1$, $Inc = 0$. An average score was then calculated for each individual. A range of .50 to 5.50 in scholarship values was obtained. The

Pearson coefficient of correlation between the scholarship values and militarism-pacifism was found to be +.15.

This low correlation seems to indicate that there is little or no relation between scholarship and militarism-pacifism. If there is any, then there is a slight tendency for pacifism to correlate positively with scholarship. Students of higher standing in scholarship tend to be pacifists.

3. *Sex.*—Among the students tested two hundred twenty-five were men and one hundred seventy-five women. The average score for men was found to be 11.54, and for women 12.30, the difference between the two being .76. The standard error of the difference was .20, which is less than a third of the difference. This indicates that women, on the whole, are more pacifistic than the men. The extent to which the two groups overlap may be seen by comparing the two distributions in Fig. 4.

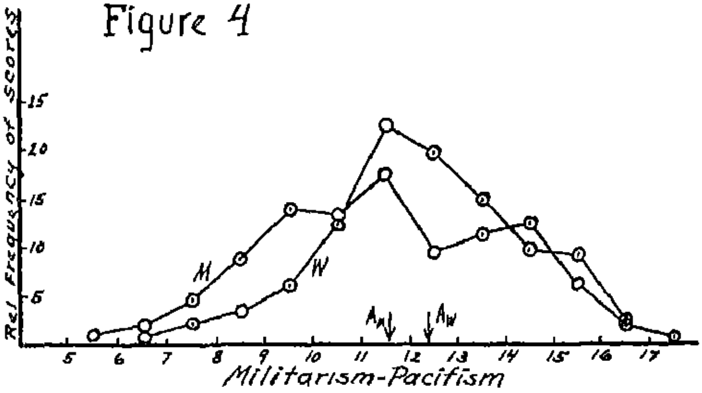


Figure 4

It is also evident from the figure that the distribution of the scores of men is wider than that of women. In terms of the standard deviations the variability was found to be 2.36 for men and 1.98 for women. Women appear to be in better agreement regarding the issue than men, and there appear to be many more militarists among men than among women.

4. *Church Affiliation.*—Data on church affiliation were obtained from the university files. Ten denominations having the largest representation were selected for study as shown in Table IV. For each an average score was calculated, and the denominations were then arranged in descending order.

TABLE IV

Denomination	Cases	Mean score
Disciples of Christ	0	13 65
Baptists	11	12 57
Jews	33	12 41
Presbyterians	24	11 95
Congregationalists	11	11 02
Methodists	24	11 49
Christian Scientists	8	11 42
Catholics	14	11 08
Lutherans	7	10 05
Episcopalians	14	10 00

The number of cases in each denominational group is not large enough to warrant any definite conclusions, but the results are interesting and suggestive. The most pacifistic churches may be found at the top of the list and the least pacifistic at the bottom. The three churches most similar in form, the Catholic, Lutheran, and the Episcopalian, seem to be similar in attitude and occupy the lowest three positions That is to say, they tend to be the most militaristic denominations of all the ten. Disciples of Christ, Baptists, and Jews stand at the top of the rank order; Presbyterians, Congregationalists, Methodists, and Christian Scientists are in the middle

CONCLUSIONS

The following brief conclusions may be drawn from the experiment described in this article:

1 By the use of the method of equal appearing intervals it was possible to construct an evenly graduated scale of militarism-pacifism The unit of measurement was one-eleventh of the range of attitude between the most extreme degree of militarism and the most extreme degree of pacifism represented in our list of statements.

2 The reliability of the scale was found to be +.90

3. Education appears to develop pacifism as measured by the scale. A consistent change toward pacifism was found for the successive undergraduate classes and the graduate classes.

4. A coefficient of correlation of + 15 was found between militarism-pacifism and scholarship.

5 Women appear to be more pacifistic, on the average, than men Men vary more than women

6. The Catholic, Lutheran, and Episcopalian seem to be, on the average, the least pacifistic churches of the ten compared. Disciples of Christ, Baptists, and Jews appear to be most favorable to the peaceful method of solving international problems.

WHAT THE THEORY OF FACTORS IS *NOT*

C SPEARMAN

University of London

Not long ago there was held a very novel and successful photographic exhibition The pictures sent in were not as usual the best, but the most instructive. Among them were those got from under-exposures and from over-exposures; those where the immersion in the developer had been too long, and where it had been too short; where the focussing had been faulty, and where the camera had been shaken, where the lens had caused flare, or the bellows had not been light-proof; where the printing paper had been stale, or not of a suitable kind. In a word, the pictures suffered from the thousand and one ills to which the photographic art is heir The great idea of the exhibitors was, of course, that their own unfortunate experiences should serve to warn others.

Now, if not too presumptuous, I would like to suggest that a somewhat similar general service might be afforded by a book which has recently come from the assembled strength of the psychological department of the University of Minnesota, as represented by Professors Paterson, Elliot, Anderson, Toops, and Heidbreder; a volume, moreover, fortified by the unstinted financial support of the National Research Council [1]

Taking first the historical aspect of this work, we are told in it that the theory of two factors was not "definitely advanced" till 1914, having previously been only "implied." Now, if such an unhappy account can befall the distinguished Minnesota syndicate, surely there must be many others who are not quite clear on the matter and would gladly be informed. In truth, the very first publication about such factors (this was ten years earlier than the date mentioned above) supplied perfectly definite observations and at the same time a mathematical criterion so exact as to be substantially that used at present. The chief final conclusion of the whole work ran (in italics) as follows:

> All branches of intellectual activity have in common one fundamental function (or group of functions), whereas the remaining or specific elements of the activity seem in every case to be wholly different from that in all the other cases [2]

Anything more precise I, for my part, could hardly formulate to this day. Further, not only were these two factors, *g* and *s*, explicitly proclaimed, even their very influences were actually discovered and measured What is still more, this original publication already drew the practical corollary (which was so brilliantly exploited the following year by Binet and Simon) that the *g* can actually be measured, this is done by throwing together any hotchpot of tests so numerous and heterogeneous that, on the whole, the influences of the *s*'s more or less perfectly neutralize each other This, vital corollary was indicated by the very title of the article, "'General Intelligence' Objectively Determined and Measured."

Incomparably more important, however, than any such mere history of the doctrine of two factors is the exposition of its essential nature According to the authors at Minnesota, the basis of the doctrine is "simple enough", their characterization runs as follows.

> It is commonly assumed that a perfect hierarchy creates strong presumption that a general factor is at work

For ventilating this view we must be more grateful to them than ever, since it not only departs from the actual state of affairs, but turns it upside down and inside out. In truth, no "assumption" or "presumption" has ever been made at all about the occurrence, or even measurability, of the *g* and *s* Instead, rigorous proof is given that in the said case of hierarchy (more strictly speaking, "equi-proportionality"), and solely in this case, the scores in the tests or other variables can *always* be divided into these two factors.[3]

What is the good of this, it may be asked? For can not the tests equally well be divided after any other fashion? The answer is to concede that, so far as the mathematical theorem goes, other modes of division—perhaps an infinite number of them—certainly are possible But yet others, and extremely important ones, are *not* possible. Thus the tests cannot possibly (where the criterion is satisfied) be so divided as to admit of any group factors (of appreciable size). Again, the tests can not possibly be so divided as to have any general factor other than our *g* Here we notice that, although the theorem does not in itself supply any evidence that the *g* deserves the title of "intelligence," still it does disprove the existence of any other and different general intelligence Of course, there may be other mental operations that anyone may like to call "intellectual." But then these can, at most, only constitute classes. They can not possibly be general in the sense of having functional unity

The further question still remains open, however, as to why we should give preference to the division into *g* and *s* over any other of the modes of division that the said mathematical theorem really does leave open A facile reply (but quite correct, so far as it goes) is that this mode of division is the simplest. But penetrating more deeply into the essential character of empirical science, we may answer with greater breadth that these two factors have hitherto shown themselves to be those divisions about which the most important statements and discoveries can be made. To mention only a few we shall see many more later on—it was in g that the inspiration really arose for all the current tests of "general ability," or even of "mental age", only in g do all these tests to this hour find any stable anchorage amid the shifting sands of the arbitary and green-table conceptions of "intelligence", and with respect to our factors alone can be measured the scientifically indispensable "probable errors"

Another object lesson is afforded by these authors at Minnesota when they embark on the troubled seas of controversy. But here, to speak with frankness, they give some occasion for surprise. They bring against the theory of two factors the following pair of objections. The one is that the inter-correlations between their tests have turned out to be low Surely we might have expected much less eminent and careful authorities than these to know at least that the theory is in no way based on the absolute magnitudes of the correlations, be these high or low, but solely on their ratios to each other. The second objection raised is that in their results the criterion of "intercolumnar correlations" did not average +1.00 But here their luck was out altogether A look at the work which introduced and demonstrated this criterion will at once show that they have used an entirely wrong formula,[4] that, moreover, they have applied this under conditions that would render even the right formula quite inapplicable,[5] and yet further that the whole criterion was from the very beginning declared to be only provisional and has now become quite obsolete [6]

To crown all, even if they had applied the proper criterion and had obtained similar results from this also (as they would actually have done, for in this case the many wrongs end by making a right) even then these results would have been no contradiction to our work For in this particular case of theirs, our work shows that the criterion ought *not* to be satisfied.

This brings us from the preceding negations of the Minnesota investigators to their positive contributions. Here the lesson at issue is the delicate one of distinguishing between *meum* and *tuum*. In the first place they lay claim to have advanced the general hypothesis of objectively determined "unique traits"; whereas it is hard to see that these are anything more than the old factors camouflaged under a new name. And in the second place (and it is the theme of the whole book) they represent themselves as having discovered the "unique trait" of "mechanical ability"; whereas in truth the latter had already been discovered (and with what I am afraid was far better evidence and more penetrating analysis) by Dr. J. Cox, who moreover succeeded in so doing by basing his research on the theory of two factors.[7]

Perhaps a clue to the last mentioned and gravest lesson supplied by these authors is to be found in their bibliography. Most concerned here is a book of the present writer, "The Abilities of Man." As was said in its foreword:

> This work is the product of many hands and much patience. The lines of investigation were suggested—and even extensive beginnings made to follow them up—over twenty years ago. Since, there have been carried out a long train of laborious researches, each bringing, as it were, a single stone upon a pre-conceived plan. And here, in this volume, at last, every stone is fitted into its place to build up the common edifice.

Among other things, the book contained a full account of this very mechanical ability at issue. Nevertheless, although this book is actually quoted in the bibliography from Minnesota yet its most relevant portion is wholly overlooked. The authors insert there the oddest lot of the same writer's minor controversial articles.

Now such action of theirs would seem to express a view that is held very widely indeed, so that once more we must be grateful for the present opportunity to comment upon it. This view is to the effect that the work of our school has consisted in arguing on behalf of some highly controversial theorem, and that the only concern of other people with it is to join in the fray, for or against.

In truth, the work of this school may be trenchantly divided into three portions. The one which should be taken first is made up of theorems which are no longer seriously contested by anybody. There is the already mentioned purely mathematical demonstration that, when the tetrad differences are zero, then (and only then) the scores of every individual in every test can be divided into g and s where these are independent of each other as also of all the other s's. Then comes the determination of the probable errors, not only of all these but also of the tetrad differences. For we require to know how large their differences ought to be when, as always happens, the data available are not the "true" correlations, but only values more or less affected by the errors of sampling Another theorem, this time obtained not from mathematics but from observation, is that under *some* conditions at any rate the observed tetrad differences do show themselves to have just this magnitude required for the two factors. So much for the portion of the work that should be taken first.

The portion that should, on the contrary, be taken last deals with the attempts to arrive at *ultimate explanations;* it includes, for example, the hypothesis that *g* represents a general "energy." Here, indeed, are problems that will long—perhaps always - be infected more than enough with "presumptions and assumptions." Nothing but trouble can be expected from even approaching them until after arraying and considering all the available relevant facts.

This brings us to the remaining and much the largest part of our work, the one which, accordingly, occupies the great bulk of "The Abilities of Man," constituting its second part, and entitled The Fundamental Facts. Here, as mentioned above, are recorded an immense mass of observations untiringly accumulated by very numerous investigators during a great number of years One large share of these observations is about the conditions under which the criterion of tetrad differences is or is not satisfied. Already in 1906 the discovery had been made that such satisfaction was liable to be disturbed by what were called "broad" group factors.[3]

These too have been made to yield both measurements and the probable errors of these. Among the more important that have (by means of the probable errors) been shown to be significant are the logical, psychological, verbal, arithmetical, and inventive abilities, as also just the mechanical ability now being put forward at Minnesota as new.

Much broader still, but no longer simply an ability, something rather that depends on ratios between abilities, have come the factors *o* and the still more wonderful *p* (sometimes written as *c*). Hardly less important have been the negative results obtained in this way, the evidence has been that the great majority of the commonly alleged unitary abilities, types, profiles, and so forth are not unitary at all.

All such factors are originally determined in an almost purely statistical manner and therefore are psychologically nearly meaningless. But then come observations that do infuse them with meaning The relations of all of them have been ascertained to all the qualitative laws of cognition; and in this way the "noegenetic" nature of *g* has been settled, not by biological or philosophical cogitation, but by actual and detailed measurements of its magnitudes and correlations under varying conditions. At the same time, investigation has been made of the relation of all these factors to the quantitative laws, to those of span, of retentivity, of fatigue, and of conation, as also to purely physiological influences. Furthermore, from all this systematic examination of the sphere of cognition, a passage has been found over to the other great sphere of the mind, that of volition, emotion, and character. Here too a general factor has been discovered, one nearly or quite independent of *g*, and perhaps even more important still.

These observations are, indeed, so presented as to throw light upon the factors involved, general, specific, and group. But they stand equally open to being envisaged from any other angle. They do not in the least depend on any such hypotheses as that of "energy," but on the contrary supply the very facts upon which such hypotheses ought to be accepted or rejected. These observations should not, I like to think, be overlooked by any one interested in any sphere at any rate of individual differences. They cannot legitimately, I feel sure, be disregarded by those who lay claim to original discoveries. Least of all ought they to be ignored by those who make bold to enter into the lists of controversy.

References and Notes

1. "Minnesota Mechanical Ability Tests." 1930.
2. "General Intelligence Objectively Determined and Measured." *American Journal of Psychology*, 1904, p. 284.
3. Positive proof will be found for instance, on pp. iii–vi of "The Abilities of Man" by the present writer, 1927. This demonstration consists throughout of straightforward mathematics; there is not an assumption or presumption from end to end. For those who find the mathematics in it difficult, it may be mentioned that a much simpler version of it has been devised and will shortly be published (in "Nature") by the well known mathematician, Professor Piaggio.
4. *British Journal of Psychology*, Vol. V, 1912, p. 82, formula (8). The simpler formula employed by them takes no account of the dominant influence of the sampling errors of the correlation.
5. *Ibid.* P. 56, the lines in italics.
6. "The Abilities of Man." By the present writer. 1927, pp. 138–140.
7. "Mechanical Aptitude." J. W. Cox, 1928, Methuen and Co.
8. Die Korrelation zwischen verschiedenen geistigen Leistungsfähigkeiten. *Zeitschrift für Psychologie*, Vol. XLIV, 1906, p. 103.

A REPLY TO SOME RECENT CRITICISMS BY PROFESSOR SPEARMAN

H. D. CARTER

University of Minnesota

Since scientific controversies frequently lead to clarification of issues, a comment on Professor Spearman's recent criticisms[1] of my attempt to analyze the nature of mechanical ability[2] may be in order.

His main criticism was directed against my use of the intercolumnar correlation method as a criterion for the presence of the familiar "two factors," *g* and *s*. He characterizes this method as obsolete, yet in his book, "The Abilities of Man," on page 139, he has a table of twenty series of data which were analyzed by this method, and the results were used in support of his theory. This suggests that even though the method may now be labelled as "obsolete," nevertheless it is permissible to continue to report results secured by that method Furthermore, it is to be remembered that the newer tetrad-difference method of analysis mathematically expresses the same fundamental relations. In my own work I recognized this fact and accordingly analyzed my data by means of both methods, discovering that the trend of results was the same. This was inevitable in the very nature of things. Since I employed the tetrad difference method in addition to the intercolumnar correlation method, I do not believe that Professor Spearman was justified in his main criticism

Mr Spearman also criticized my work by stating that my tables of correlations did not satisfy the only condition under which the intercolumnar criterion is applicable. It is true he has a criterion for the exclusion of certain pairs of columns, which is arbitrary, and not wholly free from criticism. After reading Thompson's mathematical criticisms[3] of the "exclusion criterion" I did not regard this criterion as essential, but I decided, nevertheless, to apply it to my main table of intercorrelations. It was my misfortune, however, to take this criterion directly from Professor Spearman's 1914 paper,[4] in which he himself misstated it I should have secured the method as correctly stated by Spearman in his 1912 paper[1] The criterion, therefore, was improperly applied, which led to my incorrect statement that all columns of my Table III satisfied it But recently, after reading his criticism, I applied the criterion again, this time correctly. Of the eight columns in Table III, four satisfied it, and four did not The average correlation between all columns in the table was 41. The average of the six intercorrelations of the columns which did satisfy the criterion was 30. For all other pairs of columns the average intercorrelation was 42. Correct application of the criterion is thus shown to have no important bearing on the results.

---

[1] Spearman, C A Truce to "Barking in." *Journal of Educational Psychology*, Vol. XXI, 1930, pp 110–111.

[2] Carter, H D.· The Organization of Mechanical Intelligence. *Journal of Genetic Psychology*, Vol. XXXV, 1928, pp. 270–285

[3] Thomson, G H.. On the Degree of Perfection of Hierarchical Order among Correlation Coefficients. *Biometrika*, Vol XII, 1919, pp 355–360

[4] Spearman, C : The Theory of Two Factors. *Psychological Review*, Vol XXI, 1914, pp 101–115. (See page 112)

Professor Spearman also criticized my paper because his own references to one published article and to two unpublished reports in his "The Abilities of Man" were neglected He claims that those investigations reveal mechanical ability to be unrelated to *g* and to involve a group factor But my analysis showed overlapping group factors. It is to be noted that my results differ from those quoted by Spearman in emphasizing group factors rather than a group factor. Thus it would appear that the Minnesota Mechanical Research Project makes an additional contribution to previous knowledge dealing with the abilities of man[2]

---

[1] Spearman, C. and Hart, B General Ability, Its Existence and Nature *British Journal of Psychology*, Vol. V, 1912, pp. 51–84. (See page 56)

[2] Paterson, D. G, Elliott, R M, Anderson, L Dewey, Toops, H A., and Heidbreder, E "The Minnesota Mechanical Ability Tests" University of Minnesota Press, 1930, pp. 1–610. See especially Chapter II, The Theory of Unique Traits; Chapter XI, The Organization of Mechanical Ability, Chapter XII, Mechanical Ability as a Unique Trait

STUDIES IN HANDEDNESS: III. RELATION OF HANDEDNESS TO SPEECH

R. H. OJEMANN

State University of Iowa

In this article data will be presented relative to the problem of the effect produced upon the speech function by training left-handed children to write with the right hand. For this investigation two groups of subjects were used. In the unselected group of five hundred eighteen pupils used in the investigation reported in the first article of this series of studies in handedness there were twenty-seven pupils who were diagnosed to be left-handed by the technique described in the first paper. Seven of the twenty-seven cases wrote with the right hand. These seven cases constitute the first group of dextrosinistrals[1] used in the present investigation.

The second group of dextrosinistrals was obtained as follows: In the city from which the random sample of five hundred eighteen pupils was taken from Grades III–VIII there was a total number of 1571 pupils in these grades. It was assumed that if the names of all the pupils who performed one or more of such activities as using a pair of scissors, throwing a ball, and dealing cards were obtained, several dextrosinistral cases would be secured. Accordingly, each teacher of Grades III–VIII was asked to make three lists of names: (1) A list of the names of all the pupils who used a pair of scissors with the left hand, (2) a list of the names of all the pupils who threw balls with the left hand, and (3) a list of the names of the pupils who thought they could deal cards better with the left hand than with the right hand. The purpose of this request was made clear to each teacher. It was pointed out that it was desirable to get all the pupils who used the left hand for the activities mentioned; and that, therefore, the teacher should include all doubtful cases. It was explained that it would make no difference if too many cases were reported, since each case would be carefully tested. It was also explained that it would be no reflection upon the teacher's ability if too many cases were reported. Each teacher responded enthusiastically. From the list of names submitted by the teachers the names of those students who had been included in the unselected group were checked The pupils whose names remained were carefully tested. By the use of this procedure sixteen cases[1] were secured in which the combined score on the five unimanual handedness tests was negative four or lower and in which the right hand was used for writing. These sixteen cases constitute the second group of dextrosinistrals used in the present investigation

---

[1] This term was adopted by Ballard Sinistrality and Speech, *Journal of Experimental Pedagogy*, Vol I, 1911–1912, pp 208–310, and will be used in this report to designate a left-handed individual who has learned to write with the right hand.

In addition to the unimanual and bimanual handedness tests each dextrosinistral was given three speech tests:

1. *Articulation Test.*—This test consisted of a series of sentences which were read aloud by the subject while the experimenter checked on a record sheet provided for the purpose those sounds that were made incorrectly.

2. *Spontaneous Speech Test*—The subject was shown a picture and was asked to describe aloud what he saw. The experimenter made detailed notes of any speech defects

3 *Oral Reading Test*—Gray's "Oral Reading Check Tests" were used. A record was kept of the time required to read the whole selection rather than each paragraph. This modification was made in order that the experimenter might direct his undivided attention to the errors made, since an accurate record of the errors rather than of the rate of reading was desired.

In addition to the data gathered by means of the handedness and speech tests, the parents were interviewed to obtain their judgments concerning the handedness of the subjects and to obtain the facts relative to the history of the subjects' speech and writing habits.

The data for the two groups of dextrosinistrals, twenty-three cases in all, are presented in Table I The first seven cases are the dextrosinistrals of the unselected group of five hundred eighteen pupils The cases in each group are arranged according to the grade reached in school at the time this investigation was made. Since the investigation was made shortly after the second semester had begun, the age of the subject at the beginning of the second half of the school year was used for the data in the column headed "Age" The unimanual test scores are given in terms of unit scores. (For a definition of unit scores see the first paper of this series)

---

[1] The proportion of dextrosinistrals in the unselected group is 0135 ± 003 Since sixteen cases were found in the remaining group of pupils, the proportion is 0152 ± 003. The difference between the two proportions is 0017, which is less than the probable error of either proportion

TABLE I.—DEXTROSINISTRALS STUDIED IN THIS INVESTIGATION

Subject number	Sex	Grade	Age	Unimanual test scores (unit scores)						Bimanual test scores		Speech test results	Parents' judgments of handedness
				Cut [2]	Tap [3]	Thr. [4]	B-P [5]	N-T [6]	Combined	Sw.,[7] etc	Bat		
494	G	3	9-6[1]	−5	−1	3	−1	−3	−7	L	L	No speech defects	Left-handed
184	B	5	10-8	0	−1	−3	−2	−2	−8	R	L	No speech defects	Left-handed
407	B	5	10-2	−3	0	−3	0	−3	−9	L	R	Faulty articulation	Left-handed
104	G	6	11-6	−2	0	0	−2	0	−4	L	R	No speech defects	Left-handed
19	B	7	12-6	1	−1	−3	−2	−2	−7	L	R	No speech defects	Left-handed
271	B	8	13-5	−2	0	−3	−1	−3	−9	L	L	No speech defects	Left-handed
272	B	8	13-7	−1	2	−3	0	−3	−5	L	R	No speech defects	Left-handed
735	B	3	9-2	−1	−1	−3	−1	−1	−7	L	R	No speech defects	Left-handed
734	G	3	16-6	−1	2	−3	−3	0	−5	L	R	No speech defects	Left-handed
652	B	4	9-6	−2	−2	−3	0	−1	−8	L	R	No speech defects	Left-handed
676	G	4	10-1	4	−1	−3	−6	−6	−12	R	R	No speech defects	Left-handed
609	B	5	10-9	0	−3	−3	−2	−1	−9	L	R	No speech defects	Left-handed
751	G	5	10-9	−5	−1	−3	0	−2	−11	L	L	No speech defects	Left-handed
619	G	6	11-5	−3	1	−3	−2	−3	−10	R	L	No speech defects	Left-handed
636	G	6	11-3	−2	−2	−3	−1	−2	−10	L	L	Lisp	Left-handed
703	G	6	11-2	1	−2	−3	−1	−3	−8	L	L	No speech defects	Left-handed
705	G	6	11-4	−1	0	−3	−1	0	−5	L	L	No speech defects	Left-handed
603	B	7	12-3	1	−1	−3	−5	−3	−11	L	L	No speech defects	Left-handed
692	B	7	13-6	0	−2	−3	−2	−1	−8	L	L	No speech defects	Left-handed
708	B	7	12-6	0	3	−3	−1	−6	−7	R	R	No speech defects	Left-handed
748	B	7	12-7	−4	1	−3	0	−1	−7	L	R	No speech defects	Left-handed
760	G	7	13-2	−2	−1	−3	−2	−2	−10	L	R	No speech defects	Left-handed
773	G	8	13-3	−1	−1	−3	0	0	−5	L	R	No speech defects	Left-handed

[1] To be read, nine years and 6 months

[2] Cutting.

[3] Tapping.

[4] Throwing

[5] Block-packing.

[6] Needle-threading

[7] Combined score on sweeping, raking, and shoveling

A study of the ages of the subjects and the grades reached in school shows that one of the subjects—Subject 784—was seriously retarded in her school work. This subject had an IQ of sixty-two[1] and was afflicted with a mild epileptic tendency. According to her third grade teacher, these facts accounted for her extreme retardation.

A study of the combined unit scores on the unimanual tests shows that all the cases scored negative four or lower and are, therefore, definitely left-handed. This diagnosis was corroborated in every case by the parent's judgment.

A study of the speech test results shows that in two cases, Subject 407 and Subject 636, speech defects were present at the time of the investigation. Subject 407 failed to articulate accurately the "b," "p," and "lz" sounds. His difficulty was diagnosed as faulty articulation. This subject showed no tendency to stutter. Subject 636 had a mild lisp. She substituted the "th" sound for the "s" sound. No additional speech disturbances were detected by the spontaneous speech test or by the oral reading test.

The results of the speech tests were confirmed by the interviews with the parents. According to the parents' reports only Subjects 407 and 636 had speech defects at the time of the investigation.

Subject 104, at about two years of age, when she first began to talk, stuttered noticeably. This speech disturbance continued in mild form until she reached the third grade. The disturbance then disappeared. No speech defects were present at the time this investigation was made. The first grade teacher attempted to teach the subject to use her right hand for writing and was successful without much difficulty. The mother stated that this was due to the child's temperament. The child was easily managed and the teacher, therefore, had little difficulty in starting the child to practice the use of the right hand in writing. Since the first grade, the use of the right hand for writing had been consistent. The parents reported that this training did not increase the severity of the stuttering.

Subject 10 developed a very marked case of stuttering when he was about two years of age. This disturbance lasted for about a year, after which it gradually disappeared. He had developed no speech disturbances since that time. During the pre-school period the subject used the left hand for writing. Under the influence of the first grade teacher, who treated the child's difficulties sympathetically, the subject learned to write with the right hand No effect upon the speech function of the training in the use of the right hand for writing was observed by the parents

---

[1] The writer is indebted to the school authorities for this information. The Binet-Simon test was used.

Subject 676 stuttered slightly when she first learned to talk. This disturbance had disappeared by the time the child was three years of age. No speech disturbances had developed since that time The subject used her left hand for writing until she entered the second grade Her second grade teacher persuaded her to practice writing with the right hand The teacher encouraged the child and exercised considerable patience in helping her overcome the difficulties Since the second grade, the subject has used the right hand for writing.

When Subject 703 was about three years of age the mother attempted to train her to perform the principal unimanual activities right-handed The child began to stutter The mother became alarmed and, on the advice of a neighbor, discontinued the training The stuttering disappeared about a year and a half later and did not appear again. When the child entered the first grade the parents insisted that the teacher train the child to write with the right hand. The teacher and the parents treated the child's difficulties sympathetically. No effect upon the speech function of the training in writing with the right hand was observed.

The foregoing data may be summarized thus: Two of the twenty-three dextrosinistrals evidenced speech defects at the time they were tested. In these two cases the speech disturbances appeared when the child first learned to talk. Four of the twenty-three subjects were reported to have had a speech disturbance at some time previous to the investigation. In these cases no connection between the training in using the right hand for writing and the speech disturbance could be established. The remaining seventeen dextrosinistrals had developed no speech defects at any time.

Before an interpretation is made of the above data two additional facts relative to the history of the subjects' writing habits must be noted In all twenty-three cases the intensive training in the use of the right hand was started after the speech habits had become well organized, that is, after the subject had entered the first grade. In two cases punishment was used by the teacher to train the child to write with the right hand. The second grade teacher reprimanded Subject 619 severely and punished her by slapping the hands, shaking, etc. The child had considerable difficulty in learning to write with the right hand. Her early attempts produced writing in mirror form. The treatment she received by her second grade teacher made her nervous and irritable but did not, according to the parent's report, affect her speech. Subject 692 was punished by slapping the hands and by shaking. His first grade teacher frequently strapped the left hand to the body to prevent the subject from using it during the writing period. This treatment made the child nervous but did not produce a speech disturbance.

The data presented in this paper are not subject to any serious limitations. In all twenty-three cases there is no doubt concerning the handedness of the subjects or concerning the speech disturbances. The data form an impressive array of facts the interpretation of which will be briefly considered.

These data tend to show that in training a left-handed individual to write with the right hand, the handedness of the subject is not a sufficient condition to produce a speech disturbance. It appears to be the exception rather than the rule for a speech disturbance to be produced by training left-handed individuals to write with the right hand. Any explanation for those cases in which a speech disturbance follows the shift from the left hand to the right hand must therefore be based upon a factor or group of factors which operates only in exceptional cases. An explanation which is based solely upon the neurological basis of left-handedness is therefore inadequate. The nervous system is far more flexible than is implied in such theories. The various motor and speech areas in the cortex are not localized to such an extent that it is impossible for a left-handed individual to learn to write with the right hand without disturbing the speech coordinations. Most left-handed individuals who learn to write with the right hand do not develop speech disturbances.

It is, of course, possible that in training a left-handed child to write with the right hand, factors other than the conflict between the right and left hand may be brought into play to produce a speech disturbance. For example, if the subject is emotionally unstable and is given no special care, or if the teacher employs a method of training that produces a serious mental confusion on the part of the subject, the probability of producing a speech disturbance is much greater than if an attempt is made to treat the child's difficulties intelligently.

The data presented in this study show, however, that under ordinary conditions the danger of producing a speech disturbance, after the speech habits have been formed, by training a left-handed child to write with the right hand is very slight. The nervous system is sufficiently flexible to allow an individual having a strong left-handed tendency to be trained to develop the complex coordinations required for writing with the right hand without bringing about a disturbance in a closely related series of fine coordinations such as are involved in speech.

THE CONDITIONING OF OVERT EMOTIONAL RESPONSES

HAROLD ELLIS JONES

University of California

The laboratory record which has been chosen for this report derives interest from the fact that it duplicates, in a human subject, and in a single experiment, many of the typical phenomena which have been described in the literature on conditioned reflexes in the lower animals.

The experiment was conducted in an isolated chamber, on a platform inlaid with thin brass strips, the strips were one-quarter inch in width, separated by quarter inch spaces, and connected alternately to the positive and negative primaries of a Porter inductorium. For greater quiet, a fifty dv. tuning fork was substituted for the reed make-break of the inductorium. Care was taken to adjust the secondary coil of the inductorium so as to give an electrotactual impulse restricted to a range of low intensities; the stimulus value may be described as a mildly uncomfortable "tickle" rather than as a "shock." The subject for this experiment was Robert B., a child fifteen months of age, of a markedly stolid and apathetic disposition, able to walk, and possessing a speaking vocabulary limited to fewer than five words. When he was brought to the laboratory a group of toys on the platform attracted his attention. He sat down in the middle of the platform and remained contentedly at play when left alone in this situation. Two observers, concealed behind a one-way vision screen three feet away, recorded the child's overt behavior in as great detail as was possible by a simple observational method. The inductorium, batteries, and a system of bells and buzzers, were housed in an apparatus box in a corner of the room. Robert's original reaction to an electric bell, which was sounded for approximately two seconds, could be described as "mild interest," shortly becoming "indifference." After four repetitions he no longer gave any observable reaction to this auditory stimulus. Stimulations were now given, at intervals of about fifteen seconds, associating the bell with the electrotactual stimulus. After two seconds with the bell, the inductorium circuit was established, and auditory and tactual stimuli were continued together for three seconds. The application of the tactual stimulus (felt either on hand or foot, or both) resulted in a prompt shifting of the skin areas which were in contact with the conducting strips, in each case a slight but unmistakable startle reaction was registered, accompanied by expressional changes such as a box-shaped mouth, or a puckering or pouting of the lips, and frequently by vocal murmurs or a slight whimpering which ceased at the end of the stimulation period In the periods between stimulations, Robert played quietly with the toys, and showed no overt indications of a persisting apprehensiveness or emotional upset. After three associations with the primary stimulus, the bell was sounded alone; for the ensuing five stimulations, startle reactions were recorded which were indistinguishable from those elicited by the primary stimulus; the second observer, who of course heard the bell, but had no direct means of knowing when the electrotactual stimulus was used, judged that the associations with the unconditioned stimulus were being continued through eight (instead of through merely the first three) successive trials Two stimulations by a buzzer were also effective in giving startle reactions, although with appreciably greater delay A small hand bell, of very different timbre and frequency from the conditioned stimulus, elicited no reaction. In the following trials an extinctive inhibition was established to the bell, and six reinforcements were given; in five of these, the electrotactual stimulus was received on the foot and not on the hand, and it was noted that the stimulus to the foot showed a reduced effectiveness. At the fiftieth trial, after eighteen presentations of the conditioned stimulus alone, a C-R was still present: The subject moved his foot sharply, turned his head away, and vocalized in a manner interpreted as "scolding"

After twenty-four hours, the child was again brought to the laboratory He entered the chamber readily, seated himself on the platform, and began playing with blocks At the first presentation of the bell he paused and listened, at the second he started slightly and murmured; at the third he began to whimper, and showed overt movements similar to those which had been elicited at the beginning of the experiment by the primary stimulus For the buzzer, disturbance was less marked, and for the hand bell only attention responses were recorded A small piece of *zwieback* provided Robert with an activity which for the time being inhibited overt responses to the bell. In the later stages of extinctive inhibition of the overt response, it was noted that "implicit" responses occasionally still occurred, as illustrated by pupillary dilation or by flushing.[1]

---

[1] A fuller instrumental study of conditioned reactions has been reported in Jones, H. E.: The Retention of Conditioned Emotional Reactions in Infancy *Journal of Genetic Psychology*, Vol. XXXVII, 1930, pp 485–498

After seventy-two hours, the first application of the bell resulted in momentary crying, on the second he cried and withdrew his hand slowly; on the third he showed a generalized bodily startle reaction. These were the most marked responses which were obtained during the entire series, and the experiment was discontinued in order to avoid further emotional disturbance. It should be noted that the conditioning was distinctly to the bell and not to the platform, in the intervals between stimulations he played about freely in contact with the brass strips, and showed no hesitation in touching them There was no conditioning against the total situation (of entering and playing in the room), and no indication that the experiment produced any harmful carry-over in the child's normal play activities

Opportunities were lacking to test the later course of the C-R, in view of the mildness of the primary stimuli employed in the original conditioning, it may be expected that the C-R would be short-lived, the effect of telephone and other bells would be to inhibit rather than to maintain the reaction. In conducting an experiment of this nature, it is obviously desirable to select a child of a stable make-up and to proceed with careful attention to the possibility of psychological damage. At the age under consideration, and with properly supervised conditions, it is fair to say that the experiment involves less risk of emotional upset, or of any persisting ill effects, than an equivalent amount of time spent in a routine physical examination.

For purposes of classroom presentation, the writer has found it useful to summarize the foregoing data in the form of the following table:

Stimulation number	Stimulus	Overt response	Process
	Current	+ +	Startle reaction to unconditioned stimulus
1–4	Bell	+	Investigatory reactions
5–10	Bell	−	Negative adaptation
11–13	Bell, current	+ +	Conditioning
14–18	Bell	+ +	Startle reaction to conditioned stimulus
19–20	Buzzer	+	Generalization of the C-R (irradiation of excitation)
21	Hand bell	−	Specialization of the C-R (discrimination)
22–24	Bell	+	
25–26	Bell	−	Extinctive inhibition
27–32	Bell, current	+	Reinforcement
33–50	Bell	+	Re-established C-R
After twenty-four hours			
51–54	Bell	− to +	Summation (?)
55	Buzzer	+	
56	Hand bell	−	
57–60	Bell, food	−	Temporary external inhibition
61–62	Bell	+	
63–65	Bell	−	Extinctive inhibition
After seventy-two hours			
66–70	Bell	+ +	Spontaneous recovery (disinhibition)

CERTAIN AMBIGUOUS TERMS IN EDUCATIONAL PSYCHOLOGY

STEPHEN MAXWELL COREY

De Pauw University, Greencastle, Indiana

It is difficult to understand why educational psychologists should be so little concerned with the terminological status of their field. Many words, if we may judge by the frequency with which they appear, are intended to be extremely meaningful, but what they mean is vague Yet textbook writers state dogmatically that a "habit" is this, or "learning" is that, without apparent realization of or interest in the fact that these terms mean one thing to them, and something else to the majority of their readers Nor does it seem wise to smile at this disagreement, and contend that it is after all just a matter of words. Any science must depend upon language for its development, and the value of a science is largely a function of the precision and accuracy of the words that symbolize its meanings When words are used loosely, "the mind approaches a condition where practically everything is a thing-um-bob or what-do-you-call-it" From a certain point of view, educational psychology seems to be in some such state Too many of the technically used words are indefinite They are awkward tools for thinking and treacherous as well for their ambiguity leads to confusion, and an inability to distinguish between separate factors

In an attempt to make this criticism more concrete, some of the meanings attached to the words habit, learning, intelligence and instinct are presented. These are four of the oldest, seemingly most revered, and certainly most widely used terms in educational psychology Volumes have been written concerning each, yet none conveys a precise meaning Each is understood, after a fashion, and freely interpreted All writers use the same symbols, but the ideas for which the symbols stand are practically indeterminable.

To illustrate, consider the term "learning" Probably no single, so-called technical word, is more widely used by men in the field of education. There are learning curves, learning periods, the learning process, psychologies of learning, and so on almost indefinitely To assume, however, that writers on these subjects have in mind a common idea would clearly demonstrate a lack of acquaintanceship with the facts The word "learning" is a standardized symbol representing an unstandardized concept. It is defined by educational psychologists in one or more of the five following ways:

1. Learning as growth
2. Learning as synonymous with memorizing.
3. Learning as representing any more or less permanent change in behavior not due to maturation
4. Learning as synonymous with education, or approaching a goal
5. Learning as connection-forming.

Kate Gordon is one who advocates the conception of learning as growth.[1] She writes of the growth of motor capacities and instincts. Learned behavior is behavior that has "grown" to be more elaborate and subtle Growth is always in the direction of greater complexity, so that all change is not growth, and hence not learning Learning takes place only in the transition from simple to complex. Judd, Koffka, Buhler speak of the "growth" of the mind, the "growth" of certain capacities and abilities. They are disciples of the developmental type of educational psychology Everything is growth, a rounding out, "closure."

To another group learning merely means memorizing Seashore devotes an entire chapter to what he calls the learning process, but he discusses only memory.[2] Pillsbury assumes a similar point of view when he states that "learning is no more than the formation of associations."[3] This conception of learning was quite common among all early psychologists, and was clearly expounded and developed by Ebbinghaus Recently, however, some writers have attempted to distinguish between memorizing and learning, maintaining that the former is pre-eminently a function of the cerebrum, while learning is palæencephalic.

Some psychologists, in an attempt to render an all-inclusive definition of learning, state that it is comprised of all more or less permanent modifications of behavior, save those which come wholly from maturation.[4] According to this definition, learning takes place when one loses the ability, once possessed, to recite a particular poem. Colvin writes, "The learning process may chiefly be described in its most general terms as the modifications of the reactions of an organism through experience."[1] He does not rule out physiological or maturity changes, provided they be effected or influenced by experience.

---

[1] Gordon, Kate "Educational Psychology" P 32f.

[2] Seashore, C. E.: "Introduction to Psychology." Chap. 13

[3] Pillsbury, W. B.: "Fundamentals of Psychology." P 305

[4] Cameron, E H · "Educational Psychology" P. 105 Learning in its broadest sense comprises all modifications of behavior except those which result from purely physiological changes, and those which come wholly from increased maturity

To other writers, learning has always had philosophical implications. Claparède, in a section of his book where learning and education are used interchangeably, writes, "It [?] is not merely a matter of exercising the intelligence, of furnishing the memory, but rather of directing the character upright, of stimulating zeal, and of developing the will and personality."[2] This concept of learning as a change of behavior in a certain direction, toward a goal, is rather common. Learning is called "profit from experience,"[3] it requires evaluation, and therefore cannot be studied solely with the methods at the disposal of psychologists. How can psychology decide whether a certain change in behavior represents profit or loss?

Again, learning is not only considered as a modification of behavior, but also as a modification of neural conditions. In the face of our relative ignorance of its nature, many writers in education and psychology define learning in terms of the changes in the nervous system which accompany it. Thorndike's statement that "learning is connection forming" carried weight. Starch reproduces the thought in these words: "Probably all forms of learning can be reduced to one relatively schematic type. Reception of impressions through the senses; assimilation, analysis, and combination of processes in the mind; and redirection of impulses to produce a reaction; or in brief, stimulus, association, and response."[4] Gates writes similarly, "learning consists in the strengthening and weakening of connections between situation and response."[5]

The natural result of this lack of agreement as to the nature of learning is apparent. The interested and intelligent layman thinks that educators don't know what they are talking about, and if they cannot agree among themselves as to what learning is, how can they teach his child? His reasoning is not easily criticized, and the lack of agreement is all the more confusing because it is not merely one of words, but involves ideas as well.

---

[1] Colvin, S. S. "The Learning Process." P. 1.

[2] Claparède, Ed. "Experimental Pedagogy." P. 58.

[3] Griffith, Coleman R. "General Introduction to Psychology." P. 119. In other words, true profit from experience (learning) must represent a functional change in the nervous system.

[4] Starch, Daniel. "Educational Psychology" (Revised Edition). P. 127.

[5] Gates, A. I. "Psychology for Students of Education." P. 238.

The concept of "habit" is in much the same predicament. There are almost as many different ideas as to the nature of habit as there are outstanding educators. Some attach a different meaning to the word each time it is used. Obviously, like learning, habit formation may be and is considered from two basically different points of view, neurological or behavioristic. Many authors make this distinction clearly, but in their subsequent use of the term, no hint is given as to which side of the coin is being considered. Warren defines habit as both a "process of forming connections in the nervous arc"[1] and as "an individually acquired and stereotyped series of responses or thoughts."[2] Is each of these a habit? Is a habit only present when both conditions are fulfilled? Do the two occur concomitantly, or does it make any difference?

More pertinent to the subject, even among those authors who attain and consistently hold one point view, neurological or behavioristic, there is a lack of unanimity as to what a habit is. Some, for example Judd who speaks of "Learning or habit formation,"[3] consider the term as equivalent to learning. Whatever is learned is therefore habitual. On the other hand, habit is thought of as something which is largely automatic, mechanical, stereotyped, less dependent upon consciousness. Pillsbury defines habit as a tendency, "the tendency of all movements and of mental operations which involve little or no movement to become mechanized."[4] Or, in the words of James, "An acquired habit is nothing more than a new pathway of discharge formed in the brain, by which certain incoming currents *ever after* tend to escape."[5] The main concern of this group is the determination of a basis upon which to distinguish habits from instincts.

Nor have the possibilities for a misinterpretation of the word "habit" been exhausted. Many writers cannot use the term without involving ethical considerations. He is habitually neat, or honest, or cowardly, or courteous, he has "good" or "bad" habits. According to the many followers of Thorndike a habit is a specific, automatic response to a definite situation, but the extent to which we are said to have the habit of honesty is a function of the number of *widely different* situations to which we respond in a *generally* honest manner

---

[1] Warren, Howard C., "Elements of Human Psychology." P. 253

[2] *Ibid:* P. 403.

[3] Judd, C. H "Psychology of Secondary Education" P. 54

[4] Pillsbury, W. B.· "Education as the Psychologist Sees It." P 90

[5] James, William "Psychology," Briefer Course. P 134.

There are racial habits, specific habits, general habits, mental habits, and moral habits; in terms of the nervous system or in terms of behavior To one a habit may be anything learned, to another only those reactions which are stereotyped. Probably an author is not certain just what he means himself; he may rest assured that his own meaning is seldom the one conveyed to the reader.

The concept of "instinct" is so confusing that many authorities advocate abandonment of the term in anything approaching a scientific discussion. This ambiguity is not so much a matter of the existence of instincts as it is a conflict over the best way to explain them and which reactions should be labeled instinctive. Instincts, too, may be considered from either a neurological or behavioristic point of view. All writers stress their inherited nature, but there is little agreement as to their purposiveness.[1] Some attempt to differentiate clearly between reflexes and instincts,[2] while others state that any distinction is based upon relative complexity and is made for convenience only [3] Instincts are classified in terms of neural organization, behavior, the situations stimulating them, the purposes which they serve, or the emotions accompanying them. Some writers list many,[4] others only a few [5] There is no concensus of opinion as to whether a certain reaction should still be called instinctive after it has been modified by experience, or whether it should be termed learned.[6] In one context an instinct is a specific reaction composed of a number of reflexes;[7] in another it is a quite variable pattern characterized by its purposiveness and progress toward a preconceived or unconceived goal [8] Some writers think of intelligence and instinct as being antithetical A typical quotation, though seldom appearing in so naive a form, is: "The behavior of animals is regulated by instinct, that of man by intelligence, by reason" Opposing this view, Myers writes. "I conclude then, that instincts are not, as has been generally supposed, identifiable with reflexes; nor are they, as others have urged, a *tertium quid* beside reflexes and intelligence. According to my view and my use of the word, instinct, regarded from within becomes intelligence, intelligence regarded from without becomes instinct."[1] To Myers the difference between instinct and intelligence seems to depend upon the position of the observer.

---

[1] Yet McDougall writes "Outlines of Psychology" P 71 But the great majority of all parties would agree that we may properly call instinctive those reactions of animals which seem to be purposive .

[2] Koffka, Kurt. "The Growth of the Mind" P. 90ff

[3] Cameron, E H · *Op. cit*, p 42

[4] Thorndike, E. L. "Educational Psychology" Vol I

[5] McDougall, William · "Outlines of Psychology" Chap. V.

[6] Morgan, C Lloyd · "Animal Behavior" P 71

[7] Watson, John B "Behavior" P 106. An instinct is a series of chained reflexes.

[8] Koffke, Kurt *Op. cit.*, Chap. 3

We are instinctively kind, clever, stubborn, or critical. All that we are is the sum of our instincts and experiences, an educational platitude frequently stated but as meaningless as the statement that the word "it" is nothing more that "i" plus "t." The word "instinct" is used loosely and represents no definite idea. It may mean any one of a half dozen things, yet its use continues in our scientific discussions

Lastly, the word "intelligence," though readily and freely used in technical writings may represent any one of the following concepts.

1. Ability to adapt oneself to novel situations[2]
2. Capacity for knowledge plus knowledge possessed[3]
3. Inhibition, analysis, perserverance[4]
4. Capacity to acquire capacities[5]
5. Capacity to learn or profit by experience[6]
6. Judgment[7]
7. The functioning of two factors, one general and operating in all situations, the other specific and operating in each particular situation.[8]
8. Abstract thinking.[9]

---

[1] Myers, C. S.: *British Journal of Psychology*, Vol. III, 1910, p. 218.

[2] Pintner, R.: *Journal of Educational Psychology*, Vol. XII, 1921, p. 139. I have always thought of intelligence as the ability of the individual to adapt himself adequately to relatively new situations in life.

[3] Henmon, V. A. C. *Op. cit.*, p. 201f. Intelligence involves two factors, capacity for knowledge plus knowledge possessed.

[4] Thurston, L. L. *Op. cit.*, p. 201f. Intelligence . . . contains at least three psychologically differentiable components: (1) Capacity to redefine an instinctive adjustment (2) Capacity to inhibit an instinctive adjustment (3) The volitional capacity to realize the modified instinctive behavior into overt behavior.

[5] Woodrow, Herbert. *Op. cit.*, p. 208.

[6] Dearborn, W. F. *Op. cit.*, p. 211.

[7] Binet and Simon. *Année Psychologique*, Vol. XI, 1905, p. 195. There is in intelligence, it seems to us, one fundamental organ . . . this is judgment.

[8] Spearman, C.: "The Nature of Intelligence and the Principles of Cognition." P. 4ff.

[9] Terman, L. M.: *Journal of Educational Psychology*, Vol. XII, 1921, p. 728. An individual is intelligent in proportion as he is able to carry on abstract thinking

This list could be greatly extended. Some definitions would include moral and personality traits as a part of intelligence; some would include the emotions; some would consider the word as being descriptive of behavior, some as being a static component of the "mind"[1]

The lack of concern over this state of affairs is strikingly illustrated by Dearborn, who begins his share of a symposium upon the nature of intelligence thus, "The commonly accepted definition of intelligence is . . ."[2] The very series of articles to which he was contributing brought into relief the fact that there is no "commonly accepted" definition of intelligence, nor is there much of an idea as to its nature Pressey comes out with the startling confession that, although he spends most of his time measuring intelligence, he is not much interested in its nature.

This lack of agreement over the nature of intelligence is more than one of definition. Briefly considered there are two extreme points of view held by Thorndike and Spearman respectively, with other authorities occupying the ground in between. Thorndike believes that we have as many intelligences as we face situations, while Spearman believes that his *G* operates in all reactions Quoting Thorndike: "A table of the known degrees of relationships would abundantly confirm the statement that the mind must be regarded, not as a functional unit, nor even as a collection of a few general faculties which work irrespective of particular material, but rather as a multitude of functions, each of which involves content as well as form, and so is related closely to only a few of its fellows, to the others with greater and greater degrees of remoteness"[3]

Interpreting a similar table of relationships Spearman writes: "And thus there emerges the concept of a hypothetical general and purely qualitative factor underlying all cognitive performances of any kind . . . The factor was taken, pending further information, to consist in something of the nature of an energy or power which serves in common the entire cortex (or possibly even the whole nervous system).[1]

---

[1] Yet we run across the following quotation Monroe, Walter S "Directing Learning in the High School" P 302 Some words that are used with only a general meaning in conversation and in non-technical fields, have been assigned precise meanings in the field of education Examples of such words are Standards, intelligence, . . etc.

[2] Dearborn, W F. *Journal of Educational Psychology, op cit*, p. 211

[3] Thorndike, E L., *Educational Psychology*, Vol. III, p 366

All of the available space between Thorndike and Spearman is occupied by authorities Everyone can find a definition of intelligence to suit his temperament and capacities. As Spearman has so well put it, "The reason is now evident why all search for the meaning of intelligence has, even with the greatest of modern psychologists, always ended in failure. It is simply that, in point of fact, this word in its ordinary present day usage, does not possess any definite meaning."[2]

The fault with educational writers is that too many attempt to give a pedigreed meaning to a mongrel word Long ago the chemist learned that if he wished to convey a respectable thought he must use a respectable vehicle of expression. He no longer speaks of table salt, but rather of sodium chloride. It is as impossible to extract precise, scientific meanings from words on the tongue of every novelist as it is to inject such meanings into words which already symbolize any number of different concepts.

If popular terms can't be isolated from popular meanings new ones must be manufactured, and lecturers and novelists will not make the change.

---

[1] Spearman, C.· *Op. cit.*, p 5.

[2] *Ibid.*. P 10.

THE AUTOMATIC PREDICTION OF SCHOLASTIC SUCCESS BY USING THE MULTIPLE REGRESSION TECHNIQUE WITH ELECTRIC TABULATING AND ACCOUNTING MACHINES

DAVID SEGEL

Department of Research, Long Beach City Schools

The prediction of scholastic success in general, and in the secondary school and college field in particular, is in need of devices for shortening the labor and time necessary for calculating individual prediction scores. This is true mainly because the prediction of scholastic success usually involves the use of the multiple regression equation. The work required to calculate scores after a multiple regression equation is obtained is almost prohibitive in a practical situation. It is possible, as Hull has shown,[1] to construct special machinery which will make predictions according to regression equations that have been worked out. However, the cost of constructing such machinery for this specialized use probably will prohibit its extended use for some time. In the meantime it seems desirable to adapt the use of automatic calculating machinery more generally available to the problem

We found that the tabulator and key punch units manufactured by the International Business Machine Corporation under the Hollerith Patents were machines which would perform the operations indicated by the multiple regression equations when such equations were adapted in certain ways

We will describe the use of these machines by describing their use in a particular instance.

We wished to make success predictions in the junior college field in each of the following subject groups Physical science and mathematics, English, social studies, languages, and biological science, on the basis of the number of A's and B's earned in high school, the scores in the Thurstone Psychological Examination, Edition of 1929, and the scores on the various parts of the Iowa Content Examination, all recorded at the beginning of the junior college course.

The correlation of any one of these items with college success is usually too low to be used with much efficiency in prediction (exceptions to this general statement are shown by our own data for physics and mathematics), and therefore a combination of scores or ratings into a composite rating is desirable. With a variety of correlations of the separate items with success in a college subject present, it is desirable to weight the items differently from the natural weighting. The method which gives this best weighting is the multiple regression equation technique.

---

[1] Hull, C L.: "Aptitude Testing," p 487

For the purpose of this description we shall limit ourselves to the prediction of the first two of our college subject groups. The criterion of success in the subject groups is the mark made in the subject groups the first semester in junior college The correlations between the items and the two college subject groups are given in Table I The intercorrelations between the tests are given in Table II

TABLE I —CORRELATIONS BETWEEN PREDICTION ITEMS AND CRITERIA

Criterion	Entrance units (A's & B's)	Thurstone	Iowa			
			English	Mathematics	Science	Social studies
Mathematics and physical science	130	609	720	537	573	400
English.	551	401	356	018	000	127

TABLE II —INTERCORRELATIONS OF ITEMS

	Thurstone	Iowa			
		English	Mathematics	Science	Social studies
Entrance units	261	320	004	068	242
Thurstone		560	510	111	168
Iowa English			223	377	738
Iowa mathematics				601	221
Iowa science					157

Upon the basis of these correlations the β's for the multiple regression equations were obtained.

In order to use the Hollerith card which is used in these machines it was necessary to reduce all scores to a basis of 0 to 9 inclusive, corresponding to the numbers on the Hollerith card. This was done by reducing all distribution of scores to a basis of 0 to 9 according to the schedule of Table III.

TABLE III.—SCHEDULE OF TRANSMUTED SCORES

Sigma	Transmuted score	Sigma	Transmuted score
− ∞ to −2.0	0	0 to .5	5
−2.0 to −1.5	1	.5 to 1.0	6
−1.5 to −1.0	2	1.0 to 1.5	7
−1.0 to −.5	3	1.5 to 2.0	8
−.5 to 0	4	2.0 to ∞	9

The multiple regression equation which is in general[1]

$$\overline{X}_0 = \frac{\sigma_0}{\sigma_1}\beta_{01.23\ldots n}X_1 + \frac{\sigma_0}{\sigma_2}\beta_{02.131\ldots n}X_2 + \frac{\sigma_0}{\sigma_n}\beta_{0n.1234\ldots(n-1)}\frac{\sigma_0}{\sigma_n}X_n + C$$

where

$$C = M_0 - \beta_{01.23\ldots n}\frac{\sigma_0}{\sigma_1}X_1 - \beta_{02.13\ldots n}\frac{\sigma_0}{\sigma_2}X_2 - \beta_{0n.123\ldots(n-1)}\frac{\sigma_0}{\sigma_n}X_n$$

will be simplified because all sigmas are equal since all distributions have been reduced to a sigma basis. The equation becomes

$$\overline{X}_0 = \beta_{01.23\ldots n}X_1 + \beta_{02.131\ldots n}X_2 + \beta_{0n.123\ldots(n-1)}X_n$$

C disappears because

$$M_0 - \beta_{01.23\ldots n}X_1 - \beta_{02.131\ldots n}X_2 \ldots - \beta_{0n.123\ldots(n-1)}X_n = 0$$

The β's resulting will be in the form of a decimal fraction. We have arbitrarily changed these fractions to whole numbers ranging from 0 to 3 inclusive in order to fit out tabulating equipment. The equation for physical science and mathematics is as follows.

$$X_0 = 0X_1 + 3X_2 + 3X_3 + X_4 + X_5 + 0X_6$$

where the symbols have the following meaning in transmuted score

X_1 = Number of A's + B's made in high school
X_2 = Thurstone test score
X_3 = English test score of Iowa High School Content Examination
X_4 = Mathematics test score of the Iowa High School Content Examination
X_5 = Science test score of the Iowa High School Content Examination
X_6 = Social studies test score of the Iowa High School Content Examination

---

[1] Kelley, T. L. "Statistical Method," p. 283, formula 213

The equation is reduced to

$$\overline{X}_0 = 3X_2 + 3X_3 + X_4 + X_5$$

since $0X_1$ and $0X_6$ drop out.

Because of changing fractions to whole numbers and restricting the weighting to 0 to 3 inclusive, the multiple correlation coefficient is reduced somewhat. The multiple correlation coefficient of the above equation for mathematics and science is .80 Before the reduction and restriction above mentioned, it was .81

The equation for the prediction of English was

$$\overline{X}_0 = 3X_1 + 2X_2 + X_3.$$

The multiple correlation coefficient for this equation is .62 Before the reduction and restriction was made, it was .65.

The method of using these equations with the tabulator and key punch is as follows:

A Hollerith card is punched for each student The card used is one divided into sections of three columns each. In each section the transmuted score of the test is punched in every column. A sample of the card is given here for a student who made a score of four in high school marks, a score of five in the Thurstone test, a score of six in the English part of the Iowa test, a score of five in the mathematics part of the Iowa test, a score of four in the science part of the Iowa test and a score of seven in the social studies part of the Iowa test

Cards so punched are put into the tabulator and the wiring adjusted to fit the equation For physical science and mathematics the wires will be connected as follows: Three to the Thurstone score section ($3X_2$), three to the English part of the Iowa test ($3X_3$), and one each to the mathematics part of the Iowa test (X_4) and the science part of the Iowa test (X_5). With this wiring the cards can be run through the tabulator and the weighted scores for each student will be added up and the sum printed together with the number of the student In case the tabulator is not a printer or one which adds up the different fields, the numbers can be added up mentally and the result together with the student's number written down. The composite score so obtained may be used directly or transmuted into the scholarship scale used in the institution. Where the multiple correlation coefficient is not very high the regressive effect of the equation should be accounted for in any transmutation, that is, the original spread of the scores must be restricted to fit the effect of the regression equation

Junior college success prediction																						Pupil number				High school marks			Thurstone test			Iowa high school content examination												Opr'
																																Part I			Part II			Part III			Part IV			
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	●	●	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	●	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	●	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	●	●	●	4	4	4	4	4	4	4	4	4	●	●	●	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	●	●	●	5	5	5	●	●	●	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	●	●	●	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	●	●	●	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	●
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45

1626

Long Beach City Schools

In case the multiple correlation coefficient is high the regressive effect being small can be disregarded In this way the cards can be run through the tabulator again and again until a prediction is obtained for each subject.

The method described should be of value not only where large numbers of predictions are desirable for use in individual guidance but also for purposes of homogeneous classification upon the basis of several different items. With the data punched on the cards at convenient times during the semester, the cards can be run through and a new homogeneous classification determined for various subjects within a comparatively short length of time.

A TECHNIQUE FOR EXPERIMENTATION ON GUESSING IN OBJECTIVE TESTS

LOUIS GRANICH

College of the City of New York

At a time when the pedagogical status of the true-false test still "hangs in the balance,"[1] there would seem to be a considerable field for a technique peculiarly adapted to experimentation on the element of guessing The writer's "index of guessing" is obtained by scattering, among the genuine questions of a true-false or multiple choice examination, a number of questions which involve very obscure facts, or newly coined names. As applied here, to a group of college juniors and seniors, it gave considerable evidence of its validity; it was successfully employed to reduce guessing, proving more effective than the usual instructions against guessing; and it was suggestive of a new analysis of responses to true-false questions.

PROCEDURE

A fifty item true-false test on general psychology, containing at irregular intervals ten index questions, was administered to a lecture class in social psychology. Attached to each sheet was a pre-test of twenty items of the recall, or completion type, based as nearly as possible on the same topics as the true-false test.

The instructions preceding the true-false test differed for half of the papers. The control group (Group A) received the following full instructions against guessing:

> READ THESE INSTRUCTIONS
>
> Put a circle around Right or Wrong You are not expected to recognize all of these statements *Do not guess*, because of the usual scoring system

The remaining half of the class (Group B) found these new directions on their sheets:

---

[1] Ruch, G M · "The Objective or New-type Examination" Scott, Foresman and Company, 1929, p 365 See pp 318–368 for an excellent digest of experimentation on the subject discussed here. Note that probably the best practice in true-false testing involves instructions against guessing, and scoring which corrects for guessing errors

READ THESE INSTRUCTIONS

Put a circle around Right or Wrong. You are not expected to recognize all of these statements *Do not guess*, because of the usual scoring system.

Suppose you met this question R W The Funk Charts measure the discrimination of emotions. There are no Funk Charts, and if there were you could not have heard of them To answer this question would indicate guessing on your part There are a good number of questions of this type and of similar types scattered through this test Each one will indicate guessing if answered

Therefore answer only what you know.

To obtain a random halving, the sets were arranged alternately according to the two types of directions and were dealt out thus Each student received a complete test set as he entered All directions supposed necessary were printed on the sheets, and no questions were permitted across the room, for obvious reasons.

The groups were, then, random halves of a highly selected group, each half numbering thirty-four They were expected to differ on the true-false test because of the new instructions. The pre-test was considered advisable, in order to confirm the similarity of the two groups as to distribution and total amount of information.

The method used here for establishing a control was preferred as one which resembled the procedure of an ordinary classroom test, and did not smack of experimentation to the students. Every effort was made to give the test under classroom conditions. Several other considerations pertaining to later interpretations also caused the test to be given in the form described.

The scores of the two groups on the pre-test ranged from six to twenty. With a maximum possible score of twenty, the means were 12.56 with a PE of 2.1 for B and 12.23 with a PE of 1.9 for A The groups appear similar enough to render significant any considerable differences which might be found on the true-false tests. It should be noted that for most of the mean amounts given below the deviations were large, and where differences occurred, overlapping between both groups was considerable.

Time was not called until it appeared that practically every student had finished. The students were then asked to indicate the questions on which they had guessed. Precautions were taken to secure complete cooperation. The students were addressed informally by the writer, who is not their instructor,[1] and were told that the test in no way concerned themselves, that its purpose was purely experimental. They were asked to tear off or delete their names, or any other identifications, from their papers. The point was repeated to them that in spite of their directions they certainly had guessed on some questions, probably on a large number. They were asked to cooperate in an "interesting experiment" and to indicate by a capital *G* all guessed answers. Papers were then collected.

---

[1] I am indebted to Dr Maximilian R. Schneck of the College of the City of New York for permission to administer this examination

The writer later had the opportunity to receive objections from the class against the procedure of the experiment A number of students questioned the ability of the class to indicate its own guessing No opinion was met that lack of comprehension, suspicion, or resentment had affected the frankness of the class.

SUMMARY OF RESULTS

Evidence of the cooperation of the class is found in Table I. It will be seen that in Group A, seventy-two per cent of those questions called "guessed" were actually correct, while in Group B, seventy-five per cent were correct. It may be safely assumed that the class tended to be unreserved in indicating guessing; which was the effect desired And yet both groups indicated as guessed only half of the index questions which they answered. The explanation of this apparent paradox

TABLE I.—MEAN SCORES OF GROUP A (CONTROL) AND GROUP B (NEW DIRECTIONS)[1]

	Group A (34)	Group B (34)
Pre-test scores	12 23 (PE 1 9)	12 56 (2 1)
Index questions answered	4 2 (PE 1.7)	1 9 (1 2)
Index questions indicated as guessed	2 1	1 0
Actual questions right	27 2	25 1
Actual questions wrong	6 3	5 1
Actual questions omitted	6 5	9 8
Actual questions indicated as guessed and correct	3 8	3 3
Actual questions indicated as guessed and incorrect	1 5	1 1

---

[1] Note the size of the PE's of the index distributions for both groups The difference between the pre-test means bears a ratio to its P E of about 75, making it of significance As for the index means, the ratio of the difference to its P E is about 6 5, easily demonstrating statistical significance

will appear later It is evident, however, that the student's introspection makes a poor index of guessing.

With reference to the decrease in guessing, it will be seen from Table I that omissions were increased forty-nine per cent. The number wrong was decreased nineteen per cent, the number right eight per cent and the number indicated as guessed seventeen per cent. Comparison of this last figure with the others shows that Group B was readier to admit guessing than was Group A. It should be stressed that all results here obtained are based on a single test and on a highly selected group; so that the reliability of these differences is not finally established for all groups. To judge by the mean number of index questions answered, guessing was reduced about one-half. However, it may be argued that since Group B was put on guard against these questions, the results on the index may not be valid as a measure of guessing—that the reduction was caused by the recognition of the index questions as such, and did not indicate a reduction in guessing on the body of the test. Three evidences are offered in answer:

1 Argumentative evidence which will be presented below in the analysis of student responses.

2. Each group marked as guessed almost a precise half of the index questions answered. As a matter of fact, it will appear from Table I that Group B, as compared with Group A, tended to overestimate the number of its guessed answers. This has been pointed out above Group B was therefore less aware of guessing on those index questions which it answered than was Group A; since if they were equally aware of guessing, and readier to indicate it after the test, they should have indicated a greater proportion as guessed than did Group A. When Group B, therefore, guessed on an index question, it was at least in part due to the fact that they were relatively unaware of guessing. The relative tendency against guessing carried over from the body of the test, and explains wholly or in part the reduction in the average index score

3. Fairly conclusive evidence is furnished by these figures: For Groups A and B, the ten highest and the ten lowest pre-test scores were selected, and their index scores compared, as shown in Table II. The pre-test scores were used as the criterion of achievement, since the true-false scores were subject to varying factors and were therefore unsatisfactory. It will be seen that the best and the poorest students scored about the same on the index. For Groups A and B the ten papers were then selected which had the highest and lowest

number of index questions answered, and their pre-test scores were compared, as shown in Table III It will be seen that the students who guessed most and least on the index were equal in ability. These figures indicate roughly that no significant relationship exists between knowledge of the material which has been studied and ability to discern the index questions. An additional conclusion which is bound up with this latter would have to be, that the better students do not differ from the poorer in their tendency to guess, or rather to attempt questions which are new for them. Such questions of course constitute less of sheer guessing for the superior pupils. Furthermore, the absolute number of such questions is greater for the poorer pupils.

TABLE II —MEAN SCORES OF FORTY STUDENTS SELECTED ACCORDING TO ACHIEVEMENT

	Group A		Group B	
	Highest ten	Lowest ten	Highest ten	Lowest ten
Pre-test	15 5	9 2	16	8 3
Guessed (index)	4 5	4 5	1 6	1 7
Guessed and indicated as guessed	2 6	2 3	1 1	0

TABLE III —MEAN SCORES OF FORTY STUDENTS SELECTED ACCORDING TO INDEX SCORES

	Group A		Group B	
	Highest ten	Lowest ten	Highest ten	Lowest ten
Index score	7 5	1 7	4 1	3
Index questions indicated as guessed	3 5	1 2	2 4	3
Pretest	12 4	12 0	12 4	12 4

The validity of the index, then, is apparently not affected by awareness on the part of the student or by his knowledge of the subject covered. The index does not however purport to be an absolute measure of guessing, since it contains a certain amount of recognizable material.[1] Different index scores will result from different sets of index questions. But if the items are equated as to their indicative values, they form a basis for precise comparisons, and for correlative work. They are obviously as applicable to the multiple-choice test as to the true-false.

The Index Questions

The ten index questions are given here in the order in which they appeared on the test, followed by figures representing the responses elicited by them of right or wrong, and also the number of times they were indicated as guessed.

It will be seen that the guessing element in these questions varies from an almost absurd amount in several, to so small an amount in others as to render them almost conventional. A better index would be composed of equated items. It is not sheer guessing that is measured by most of those here, but the tendency to attack material that is new, and offers no legitimate basis for an answer.

---

[1] Nor does it measure those questions which are answered with an even chance of success If it did, the following formula could be applied:

$$\frac{N}{I - N} \times O = G$$

where G is the number of questions guessed, I the number of questions composing the index, N the number of these questions answered, and O the number of omissions on the body of the test Half of G would represent the number right or the number wrong through guessing If the formula be applied to our data above, a striking similarity is found between both groups, in their estimated "true" scores The number right or wrong through guessing is 2.4 for Group A and 1.2 for Group B Subtracting these amounts from the number right to find the "true" score, we obtain 24.8 for A and 23.9 for B. Subtracting from the number wrong to find the actual amount wrong for other reasons than guessing, we obtain for both halves an identical remainder of 3.0. This is what we should expect for such similar groups Furthermore, the ratio of their G's is about equal to the ratio of their index scores The explanation of these agreements, however, presents many difficulties They are probably spurious, unless by accident the index sampled guessing correctly. The possibility of using the index as a refinement in scoring, even for groups as a whole, requires further demonstration

It may be argued that the index questions, if equated, contain a constant proportion of guessing, and can be converted to pure guessing scores by multiplication As will appear later, however, what we encounter on the test is not pure guessing, but guessing dispersed as an element among many questions, and subject to uncertain rules of probability.

The questions are composed either of obscure facts or of propositions involving a newly coined term, and should be conventional in structure The practical application of the index and of the new instructions, remains to be further demonstrated. It may be that the present standard instructions (*Do not guess unless you have some basis for an answer*) will offer the best results, and that guessing need not be further reduced. It may be that R-W scoring is more valid for the very reason that it takes into account the wrong answers. It will prove too difficult to construct tests with reliable individual indices, and too cumbersome to score such tests, to allow of their widespread use The use of meaningless material may prove objectionable in classroom practice since it renders possible retention of matter which is undesirable. It may also be said that the new instructions are threatening and punitive in attitude. But they are no more so essentially than instructions which warn of R-W scoring; in fact they offer more reasonable justification to the student for penalizing. Whatever pedagogical faults may appear against the technique, there remain many possibilities for its use in objective tests outside of the

TABLE IV —SHOWING THE EFFECTIVENESS OF THE VARIOUS INDEX QUESTIONS

	Marked right	Marked wrong	Indicated as guessed
1 Convergence of factors will probably explain the transfer of training .	14	12	9
2 The Juke's history shows that fourteen per cent did not become public charges .	22	10	14
3 The acquisitive instinct is today considered conductive as well as dynamic	10	8	8
4 One cause of illusions is the functioning of diffused association-arcs	23	1	12
5. There are never more than three thalamic membranes	4	4	7
6 Glandular misfunctioning is the cause of pitrinosis .	9	0	6
7 Dividing a learning curve into zones increases reliability but lowers validity .	10	10	11
8 There are more sensitized pigment spots in the brain than in the spinal cord	17	7	10
9 Laselles found the white and colored races about equal in keenness of perception	17	11	18
10. Jung was the first psycho-analyst to refute the existence of the negative personality	9	9	9

classroom. It will be noted that the possibility of detection constitutes adequate motivation for reducing guessing.

The distribution of index scores (Table V) shows that extreme amounts of guessing are almost entirely eliminated, while there is a much greater frequency of very small amounts.

The writer found these questions easier to prepare than those which were real. Care must be taken only that obscure facts be obscure enough, and that an answer of *False* should not by some chance be an honest and correct answer to a sentence involving a neologism, as for example, in "The acephelon is located in front of the thalamus."

The greatest utility of the technique lies in further theoretical investigation. Among the possibilities are: (1) Correlative work and the determination of relations between guessing and the different elements of a test situation. (2) Comparison of the amount of guessing going into different types of tests, including even completion tests, also, estimation of the relative amount of guessing going into specific tests. (3) A study of the distribution of student-types as regards guessing. (4) Analysis of characteristics of questions which invite guessing. (5) Study of the extent to which students are aware of their guessing. (6) Retention of content with different tests. (7) A measure of one phase of classroom discipline, and of the effectiveness of different instructions, motivations, personalities, etc. (8) Experimentation leading to the perfection of the index, confirming its validity, and determining its possible application to the classroom. The possibility of employing it as a refinement in scoring is of greatest importance here.

The dilemma appearing from our results is this: While only half of those index questions answered are indicated as guessed, three-quarters of those so indicated on the test proper are correct. It is of course assumed that the effect of "specific determiners" alone will not explain this. The analysis which will presently be given is suggested in explanation of this phenomenon. It does not, of course, derive any definite corroboration from our figures, but is merely suggested by them.

TABLE V.—DISTRIBUTION OF FREQUENCIES OF GUESSING

Index score	0	1	2	3	4	5	6	7	8	9	10
Frequency, Group A	1	3	5	5	10	2	0	3	2	2	1
Frequency, Group B	8	10	6	6	1	2	0	0	0	1	0

THE NATURE OF RESPONSES

Discussion to date has assumed that the response to any question is determined by one of several factors; either information, guessing, misinformation, or a "hunch" resulting from the characteristics of the question (Wiedemann's *specific determiners*). It has been recognized that guessing varies at different times, with different groups, under different test conditions, and also with each individual examinee It has been found that guessed answers are more often correct than incorrect;[1] and this fact has been explained on the basis of subliminal configurations of knowledge. There is also evidence that R-W scoring over-penalizes students;[2] and it has been pointed out, in partial explanation of this fact, that answers wrong through misinformation are penalized doubly It is obvious that this routine analysis will not suffice to explain our results in this study. A closer analysis of the factors determining a response is therefore relevant.

Questions are *factual* insofar as they measure the recognition of terms or of facts. Questions are *deliberative*, involving *judgment*, to the extent that they require the recall of facts to decide the truth of a statement not already taught A question intended as factual may sometimes be attacked in deliberative fashion, to substitute for or to supplement recognition. "Deliberation" involves recall, comparison, criticism, ingenuity, application, practice in problem attack, and any number of phases of general intelligence It may be an acceptable compensation for inability to recognize, if by its means the correct answer is arrived at, from true premises and by a correct process of reasoning.

The answer to any question is the result of a number of factors, not of one. Thus in the (rather poor) question, *Wundt established the first psychological laboratory in 1879*, a student may recognize the name *Wundt* (information) but be uncertain as to whether this man was the pioneer in question (guess). He has an idea that Titchener was the founder (misinformation) He does not recognize 1879, but decides its probable accuracy by checking it with other dates (deliberation, as well as guess) The probable accuracy of the date, and other considerations, now lead him to believe the entire statement true (hunch, or suggestibility). He answers in spite of his uncertainty (mental set as to guessing from directions).

---

[1] West, P. V A Critical Study of the Right Minus Wrong Method *Journal of Educational Research*, Vol. VIII, 1929, pp 1–9

[2] Ruch, G. M *Op. cit.*, p. 352

We should then regard guessing as a *factor* insofar as a question is answered with uncertainty, while misinformation becomes an inadequate term. There may be misinformation, it is true, as to a part of a question; problematic or partly problematic questions answered with certainty but incorrectly are fairly frequent. But practically all instances of misinformation, as defined hitherto, are accompanied more or less by uncertainty It is difficult to assume that students learn complex facts in a form precisely incorrect.

If the final score formula for the number right can not be expressed in terms of whole questions, does it tend to be the sum total of the *fractions* of information plus one-half the total of *fractions* of guessing-factors, plus approximately one-half of the *fractions* of suggestibility factors? Or does the final score depend on the proportions in which these factors are combined? Does each sentence bear its own rules of probability? (Note that the recognition of a single term in a statement may furnish a clue to the correct answer, for example, by elimination.) Do intelligence, training in problems, and application enter also as factors compensating, legitimately or not, for lack of knowledge? Is much recognizable material necessary to assure a student that he is not guessing? Above all, what rules of probability govern questions made up of different recognizable, unrecognizable, and problematic proportions?[1] Would these rules explain the over-penalization of R-W scoring? Further study might well be directed along these lines.

It need not be pointed out that many questions, especially shorter ones, are entirely recognizable or unrecognizable. It is also true that there are degrees of learning; that some facts are on the borderline of memory. This situation introduces other factors which affect scores. application, amount of time, and accident in striking upon effective associations These last few items will be recognized as universal causes of unreliability, together with such elements as copying, ambiguity of the questions, attitude of student, conditions of the test, obscure wording, etc. In problematic questions an analogue to guessing is found in the process of hitting upon a line of reasoning which leads to a correct answer, although the background of information is very limited, and the judgment poor.

---

[1] Wood found relationships between the problematic quality of questions and their reliability, as well as one aspect of their validity.

Wood, Ben D. Studies of Achievement Tests. *Journal of Educational Psychology*, Vol XVIII, 1926, pp 10 and 13.

At any rate, we cannot regard some one factor as the single cause of a given response, nor can we consider such responses at all analogous to coin-flipping. An exception may perhaps be made for simple, short, factual questions

We may now be able to explain the peculiarities of true-false tests which have already been mentioned We know, for example, that enough uncertainty may enter into a response for the student to call it guessing And yet such questions may contain enough recognizable material to be answered correctly in more than half the cases. We may now reject the dubious concept of misinformation accompanied by certainty. We may also explain why answers which are changed by the student are correctly changed most of the time.[1] We may deduce from our theory such very plausible propositions as these: A multiple choice of four answers may really offer only two actual alternatives, since two others may be conspicuously absurd Or, if one of the answers is known to be incorrect, the chance of guessing correctly will be one out of three. If, on the other hand, none of the alternatives are recognized, but one of them offers a "catch," then the chances of correctness are less than one out of four

It is obvious that if a student does not recognize a fact, he can not remember whether that fact was ever discussed in class, and can not decide whether he is expected to know it Unless the question is altogether absurd or unconventional in appearance, he can not be sure whether it is an index item, or one which he should know It is for this reason that knowledge of the presence of the index does not affect its validity This analytical evidence the writer has already promised to present. The effectiveness of the new directions is due to the same situation To guess on an item unrecognized in part is risking an increase in index score

SUMMARY

1. The index offers a valid measure, although not necessarily absolute, of guessing on objective tests. Guessing is here defined as the tendency to answer questions which are unrecognized either wholly or in part, when an answer can not be deduced with certainty from such information as the student possesses

---

[1] Mathews, C. O Erroneous First Impressions on Objective Tests Also, Lowe, M L and Crawford, C C First Impression vs Second Thought in True-false Tests. *Journal of Educational Psychology*, Vol XX, 1929, p 195 and p. 286 respectively

2 Warning as to the presence of index questions in the test used here proved to be more effective in reducing guessing than the usual full instructions and warning of penalization.

3. For a number of reasons already explained, students can not reliably indicate their own guessed answers. This fact is demonstrated by comparison of index scores with our other numerical results.

4. In this study, superior students did not differ from poorer students in their tendency to guess; that is, in their average index scores It follows that they were no better able to discern index items even when aware of their presence The students who guessed most and those who guessed least, according to their index scores, were nevertheless equal in achievement according to a pre-test.

5 Some of the results of this study were difficult of explanation in terms of the conventional analysis of student responses to true-false questions. A revised analysis is presented, in terms of which a plausible explanation can be found for many of the peculiarities of true-false tests, besides those appearing in this study.

6. Argumentative evidence, based on this analysis, is produced to show that responses are more often the result of several psychological factors than of one. Simple factual questions are an exception to this rule.

7. Furthermore, such answers are not subject to simple mathematical rules of probability in scoring; their probabilities of correctness vary according to the different proportions of guessing, information, etc, that enter into the responses.

8. Further studies employing the index technique are suggested above, and studies leading to the criticism and improvement of the index will be of value.

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

CONDUCTED BY FRANCES M. FOSTER

The Science of Psychology, An Introductory Study, by Raymond Holder Wheeler. New York. Thomas Y Crowell Co, 1929. Pp. XI + 556.

In "The Science of Psychology" Professor Wheeler of the University of Kansas makes a pioneering attempt to present general psychology in textbook form from the Gestalt point of view In spite of the difficulty of the task, the result is neither one-sided nor incoherent Nor is it (perhaps more surprisingly) unduly abstruse or difficult. In parts the discussion involves certain inherent difficulties, but the style is simple and the presentation logical and clear.

Whereas most textbook writers seem to base their eclecticism upon the principle that those explanations which are most widely accepted at the time of writing are to be preferred, Wheeler seems to base his upon the principle that those explanations which appear to him simplest and most inclusive, are to be preferred. It would be more accurate to say that in both Wheeler's book and in the usual textbook there is a mixture of these two principles, but that in the latter there is more evidence of the principle of general acceptance than of the principle of "parsimony," whereas in Wheeler's book the reverse is the case.

All living activity is described by Wheeler as goal activity, governed by the "Law of Configurations" and the "Law of Least Action" A goal he describes as

> the termination or end of a given response, however simple or complex . Represented by those stimuli or objects in a stimulus-pattern, the reaching of which relieves tension in the organism. Represented in the organism by a relative equilibrium in the neuromuscular system In bringing behavior into relation with dynamical principles, the goal is interpreted as a point in an energy system toward which a body moves under stress, it is the point of low stress of the system, and the organism is regarded as part of the system in which it moves The situation is akin to a gravitation system with its center constantly moving, along with shifts in the alignment of potentials elsewhere in the system.

Wheeler's description of the "Goal" concept is more definite and concrete than McDougall's description of the concept of purpose and less restricted in scope than the concept of desires proposed by the psychoanalysts Although he employs it as an explanatory principle quite as universally as these other psychologists, he does not give the impression of distorting the facts to prove his point of view. On the contrary, the large amount of experimental data which he brings together in the 556 pages of his book is clarified and unified by consistent examination from his defined viewpoint.

Wheeler progresses systematically from the sociological aspects of psychology to the biological aspects, with the chief emphasis, where it belongs, midway between these two extremes, on "Emotive Behavior" and "Learning Behavior," the treatment of the latter being so particularly thorough and stimulating as to make the book an excellent reference or text in advanced courses in educational psychology.

His treatment, in the first chapters of the book, of "Social Behavior and Its Conditions" is without distinction—just another compilation, not unusually coherent, of the facts and theories generally reported in social psychologies. It has the merit of being thoroughly modern and of containing some valuable and original formulations of principles. The latter give the impression, however, of being merely inserted for good *measure without affecting the choice of other material in this* section of the book

But this deficiency in the first part of the book is in sharp contrast with the rest of the book, and since the first part is fairly typical of that vast number of texts whose chief reason for existence is their clarity and modernity, it forms a handy reference point by which to judge the superiority of the body of the work. JOHN N WASHBURNE.

Syracuse University

Manual for Determining the Equivalence of Mental Ages Obtained from Group Intelligence Tests, by Ross O. Runnels. Test Method Helps, No. II. Yonkers-on-Hudson, N Y.: World Book Co., 1930. Pp 14, paper.

Expressing Educational Measures as Percentile Ranks, by F C. and O K Buros. Test Method Helps, No. III. Yonkers-on-Hudson, N. Y.. World Book Co , 1930. Pp. 27, paper.

It has long been noted by users of tests that there are gross differences in mental ages and intelligence quotients obtained from different group tests Due to variations in methods of standardization and differences between the groups by which the various tests were standardized, a sad lack of equivalence exists between the tables of norms. Dr. Runnels administered the National, Haggerty, Otis-advanced, Terman Group and Otis Self-administering Tests to 1422 pupils in Grades V to IX inclusive The equivalence of scores was determined by equating percentile ranks For each percentile rank, the average of the mental ages given by the manuals for the various tests was taken as the new mental age norm. Of course, no such procedure will increase the reliability of a group test, but Runnels' tables will be of great assistance in obtaining more comparable results when a variety of group tests is used in a school in successive years.

"Test Method Helps, No II" gives an admirably clear and simple explanation of ranks and percentile ranks and includes tables to aid the computation of percentiles when the number of cases is from 11 to 100. For those who are not adept in the use of a slide rule, this little book provides a quicker and easier method of computing percentile ranks than has hitherto been available

LAURANCE F. SHAFFER

Carnegie Institute of Technology, Pittsburgh

Child Adjustment, by Annie Dolman Inskeep. New York, D. Appleton and Co , 1930 Pp XVI + 247

A study of the facts and principles presented in "Child Adjustment" will enable the teacher and educational specialist to gain new insight into the educational process as related to the growth process in the school child The book serves to bring together the major findings of many scattered bits of research As the author states in the preface, the essential purpose of the book is that of "finding out how the child's body, mind, and emotions differ from an adult's, how they develop into the adult stage, and how they should be cared for during school years." The topics discussed include adjustment with reference to height and weight, growth and functions of the larger muscles, handedness and footedness, the internal organs, different developmental ages, the teeth, adjustment in relation to growth of brain and nerves, the eyes and ears, adjustment in relation to growth, development of the mind, measurement of intelligence, critical adjustment periods, emotional health and mental hygiene In the appendix there is a discussion of different psychological theories, behaviorism, gestalt psychology, Freud, Jung, Adler, psychoanalysis.

The author shows originality in the methods devised for studying different factors in child behavior and makes the material of practical usefulness through her descriptions of actual school practice in studying the physiological, mental and emotional development of pupils. The best sources of scientific data have been consulted and are referred to in each chapter The interrelation of physical and mental factors in the child's development is emphasized with each topic discussed. The belief is expressed that if the child's home and school adjustments have been made with his physical, mental and emotional needs in view, the child makes better progress in his school work Cases of discipline, failure, nervous break-down, may be attributed to failure to understand the child as different from the adult.

In occasional instances the reader may feel that the author lays too much stress on physical deficiencies as major causes of maladjustment, as for example when lying and stealing are found in conjunction with adenoids or poor eyesight, and the behavior defects are attributed to the physical ones. A tendency to attribute school failure to physical defects is implied in the author's discussion, but there is also emphasis throughout the book on the importance of investigating all causal factors in an adjustment problem GERTRUDE HILDRETH.

The Lincoln School of Teachers College

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII March, 1931 Number 3

THORNDIKE'S C.A.V.D. IS FULL OF *G*

KARL J HOLZINGER

University of Chicago

The chief point of this note is to show that Professor Thorndike's well-known C.A.V D. Intelligence Test may be thought of as saturated with Professor Spearman's *g* rather than with a number of group factors. Professor Thorndike's interpretation of three Spearman laws of cognition will also be touched upon This second point is doubtless a very dangerous one because the writer is not a psychologist and has only a moderate *g*, while Professor Thorndike is not only a pure psychologist, but must have a *g* that approaches the colossal.

The data for the first point are furnished in Professor Thorndike's Measurement of Intelligence.[1] Here the C.A.V.D. test is compared with other intelligence examinations designated as follows·

C A V D . . .	X_1
Otis Self-administering Test	X_2
Terman Group Test (Grades VII–XII)..	X_3
Stanford Binet (MA) .	X_4

The intercorrelations of these four tests are given by Professor Thorndike and may be written in the form

	X_1	X_2	X_3	X_4
X_1		87	.94	78
X_2			88	77
X_3				.77

The tetrad differences[2] work out

$$t_{1234} = -.054 \pm .017, t_{1243} = -.017 \pm .010, t_{1342} = +.037 \pm .022$$

---

[1] Thorndike, E. L. "The Measurement of Intelligence" Bureau of Publications, Teachers College, Columbia University, New York City. P 96ff

[2] Professor Spearman's "Abilities of Man" The Macmillan Company

These tetrads may be regarded as statistically insignificant showing the existence of a common factor g and four uncorrelated specific factors s_i as indicated by the equations (1)

$$\left.\begin{aligned} x_1 &= m_1g + n_1s_1 \\ x_2 &= m_2g + n_2s_2 \\ x_3 &= m_3g + n_3s_3 \\ x_4 &= m_1g + n_4s_4 \end{aligned}\right\} \qquad (1)$$

The quantities m_i and n_i $(i = 1, 2, 3, 4)$ being constants.

The existence of g being indicated it is now possible to obtain its correlation with each of the four tests (see Spearman Appendix). These correlations are as follows:

$$r_{1g} = .960, r_{2g} = .921, r_{3g} = .960, r_{4g} = .817$$

If the probable error of these coefficients is of the order .01 or .02 it thus appears that C.A.V D. and the Terman test are most highly saturated with g while Stanford-Binet is the least effective measure of this factor

In his own analysis of the above correlations Professor Thorndike points out (p 96, *op. cit.*) that "Intellect C.A V.D. is very much the same as that which is measured by representative examinations for so-called intelligence" Professor Spearman's method of correlation with g not only shows that these four tests are measuring the same "intellect" but shows *how well* each measures this "intellect" and indicates that the "intellect" in common may be regarded as g as given in factor pattern (1). This seems to us a great advance over the usual crude methods of validation by correlations. The criterion by the Spearman method is a well defined g instead of a subjectively labeled test score.

It is, of course, possible to interpret the same set of correlations in an infinite number of ways and to employ an infinite number of factor patterns other than (1) A certain western psychologist prefers to think of the variables as made up of an infinite number of independent elements even when the tetrads vanish. This is a possible interpretation, but since it cannot be verified statistically and could be made regardless of the relation between correlation coefficients, it seems to us to have little scientific value.

When the tetrad differences do not vanish, group factors are presented and more elaborate factor patterns may be employed. Illustrations of this sort are given by Professor Kelley[1] and by Professor

---

[1] Kelley, T. L: "Crossroads in the Mind of Man" Stanford University Press.

Spearman (*loc. cit.*). Some of the difficulties arising from the use of such elaborate patterns are the complexity or total lack of adequate probable errors to test the theory, and the great difficulty in attaching meaning to the numerous factors employed.

By way of illustrating these points, we may take an example from Professor Kelley's book (p. 97ff.). The four tests used are:

X_1 = Reading speed
X_2 = Arithmetic power.
X_3 = Memory for words.
X_4 = Memory for meaningful symbols

The intercorrelation and tetrad differences are as taken from a paper by the writer [1]

	X_1	X_2	X_3	X_4
X_1		0580	1050	2969
X_2			.1487	2480
X_3			.	0693

$t_{1234} = -010 \pm 037$
$t_{1243} = -005 \pm 037$
$t_{1342} = .005 \pm .010$

From the insignificance of these tetrads we may conclude that factor pattern (1) with only *g* common is sufficiently complex. If group factors are present in these four tests their effect is insignificant, yet Professor Kelley employs the pattern given by the following portion of his Table XII (*op cit*). Numbers in the table are standard deviations

Tests	α = heterogeneity, maturity, sex, race	β = verbal factor	γ = number factor	δ = memory factor	ϵ = spatial factor	ζ = speed factor	Specific	
							Not chance	Chance
1 Reading speed	10	60			00	38	30	28
2 Arithmetic power	21	.	63		31		16	66
3 Memory for words	60	00	.	50			36	33
4 Memory for meaningful symbols	.50			62	30		32	30

---

[1] Holzinger, K. J. On Tetrad Differences of Overlapping Variables *Journal of Educational Psychology*, Feb., 1920

The above example is but a fragment of Professor Kelley's work on these data and is not included by way of criticism, but merely because the numeral work was at hand. As far as these four tests are concerned we hold that pattern (1) is adequate Professor Kelley employs the elaborate pattern in the above table and interprets the common factor α as "heterogeneity, maturity, sex, and race" We argue that the factors α, β, γ, δ, ϵ, ζ, etc. are insignificant in these four tests, and that whatever common factor is found may be regarded as g

At a meeting of distinguished psychologists last year several expressed grave doubt as to the meaning of g, they preferred α, β, γ, δ, ϵ and ζ probably because specific labels had been attached to them. The writer is certainly in doubt as to the correct interpretation of g, but this doubt rises to complete confusion when a half dozen other factors are added even though they are shown to be significant. We can at least get the correlation between g and other variables when the tetrads vanish and thus approach its meaning. The bases for determining the number, numerical value, and meaning of other general factors β, γ, δ, ϵ, ζ etc are much more subjective. Such factors may need to be added in the analysis of variables, but only when pattern (1) doesn't hold as shown by the tetrads. If they are added their meaning should be approached with even greater care than the meaning of g.

The whole point of their digression is that when we do get the tetrads to vanish and thus establish the adequacy of pattern (1) we should be very happy about it We have arrived at a simple and a very beautiful statistical explanation. We may also, by the Spearman technique, build up a pool of tests concentrated so highly with g that we will come to know its meaning more clearly than that of tests subjectively labelled.

It thus appears that Professor Thorndike should be pleased with our findings. If he ever makes another test more replete with g he should be still happier. He should not try to think of the common factor as "heterogeneity, maturity, sex and race," but rather have these all eliminated and show that g is still there.

Turning next to the interpretation of Spearman's laws of cognition we may test these as follows: (1) Any lived experience tends to evoke immediately a knowing of its character and experience; (2) the mental presentation of any two characters (simple or complex) tends to evoke immediately a knowing of the relations between them. This law may be termed *Eduction of Relations,* (3) the presenting of any character together with any relation tends to evoke immediately a knowing of the correlative character. This law may be called *Eduction of Correlates*. The word "immediate" here indicates the absence of intervening processes.

According to Professor Spearman, intelligence includes all processes derived from these three principles comprising the ability to apprehend experience, educe relations and educe correlates.

The second of the above laws may be illustrated by an analogies test.

Bat. Ball. Hammer: Boy: Nail Handle: Axe Here the characters or fundaments are the seven items presented The subject may educe various relations between these fundaments such as:

Bat *is used for hitting* ball.
Bat *has same initial letter as* ball
Hammer *has* handle
Hammer *has same initial letter as* handle
Hammer *is used for hitting* nail

In addition to these the subject may educe certain relations between the relations listed above. These might include:

is used for hitting is unlike *has*
is used for hitting is same as *is used for hitting*
has same initial letter as is same as *has same initial letter as*

The correct answer to the original test is "nail" arrived at by educing some or all of the above relations. The question then remains what to do with the answer "handle" selected on the basis of the same initial letter as "hammer." It may be argued that both answers are correct or that "nail" is a *better* answer than "handle" and should receive all or at least more credit than the latter. If the second argument is followed, then the subject might educe a relation of the sort.

"Use of an object *is generally more important* than the initial letter of the object"

In any case the above example appears to involve only the eduction of simple relations like the first five and more complex relations like the last which are dependent upon the first.

Professor Thorndike[1] has commented upon these three laws as follows:

> There is no doubt that the appreciation and management of relations is a very important feature of intellect, by any reasonable definition thereof Yet it seems hazardous and undesirable to assume that the perception and use of relations is all of intellect In practice, tests in paragraph reading, in information, and in range

---

[1] *Loc cit*, Pp. 19–20

of vocabulary, seem to signify intellect almost as well as opposites and mixed relations tests In theory, analysis (choosing suitable elements or aspects or relations), and organizing (managing many associative trends so that each is given due weight in view of the purpose of one's thought), seem to be as deserving of consideration as the perception and use of relations Moreover, I fear that in all four cases we need other valuations to decide which are *better* relations, or *more abstract* relations, or *more essential* elements, or the *more sagacious* relations, or the *more consistent* organization, or the *more desirable* balance of weights, and the like (Italics are Thorndike's)

To the unpsychological mind of the writer the "other valuations" cited above are largely, if not entirely, eductions of relations between relations Thus the correct solution of the Bat: Ball problem probably rests chiefly upon eduction of the sort,

use of an object *is more important than* the initial letter of the object.

This seems to us just as much an eduction as the case,

Bat *is used for hitting* ball

In fact nearly all of the "evaluation" cited by Professor Thorndike appears to us as Spearman "eduction."

As a further illustration one may take Professor Thorndike's Task 1 on p. 163 (*op cit.*) The subject is to make a true and sensible statement, filling one word in each of the blank spaces: "The . . . way to . is by airplane." The filling of these spaces appears to involve eduction of relations between fundaments of a sort. Thus the second space might be filled with such words as "travel, die, swim," etc Likewise the whole new phrase "way to travel is by airplane," or "way to swim is by airplane" may be treated as a new fundament and the first place filled by eduction resulting in words like "best, fastest, cheapest," etc. Furthermore the completed series may be related to the fundament "previously experienced sensible fact," and several such completed series related by the eduction "more sensible" Thus the subject may arrive at the sentences:

1 The quickest way to die is by airplane
2 The cheapest way to travel is by airplane.
3. The quickest way to travel is by airplane.
4. The slowest way to swim is by airplane.

The choice of the "most sensible" completion might be obtained by a series of eductions between the above completions. According to Professor Thorndike this choice is a matter of "decisions" as to "most sagacious" relations, etc. How the subject could make the "most sagacious" selection without eduction we do not see, and if all such decisions, valuations and sagaciousness are ruled out of Spearman's laws he has instead of Laws of Intelligence anything but that.

TETRAD-DIFFERENCES FOR NON-VERBAL SUBTESTS*

WILLIAM STEPHENSON

University College, London

INTRODUCTION

The work to be described constitutes an introduction to intensive research on questions of verbality and thought; but the present series of papers concerns primarily the satisfaction of the theory of factors by data obtained by use of verbal and non-verbal subtests. The question of the factor content of certain non-verbal subtests is of first importance, and is our immediate concern; later, we shall examine comparable data for a set of verbal subtests, and ultimately we shall use the one kind of subtests as "reference values"[7] for the other.

By verbal subtests we refer to g-tests involving words, or phrases or other complex linguistic structures as fundaments; the non-verbal g-tests are set in spatial, perceptual, or pictorial or other ostensibly non-verbal fundaments.

The recent researches of Line[6] and Fortes[3] are indication of the present detailed attention given to non-verbal subtests. The interest in such subtests follows from the theory (and obtained facts) of the universality of the Spearman g-factor, that which appears to characterise eductive processes[7] no matter under what conditions they function. The work of Davey[2] is recalled, where the g-factor was found to cover pictorial as well as verbal subtests. But, the correlational data gathered by Davey, myself,[10] Line, and Fortes, are for populations of the order 100 only and, to have a more secure foundation of theory and supporting fact, it seems that some data should be got from large populations, of the order 1000 at least. The first object of the present paper is to provide data gathered from a population of 1037 girls

A correlation table for one thousand population serves at least two points of value We examine the table by means of the Spearman tetrad technique. In the first place, the Spearman Theory of Two Additive Factors requires a fit in the tetrads to within sampling error value, and, instead of sampling error of amount 0 03 for one hundred populations, we should encounter sampling error of 0.0095 for one thousand population (with subtests of the kind used by us). Secondly, it is likely that any such Theory, dependent as it is upon the use of many subtests, involving scores of test items, and for data gathered under complex experimental conditions, will need to take account of errors other than sampling error. Particularly will this be so for large populations, for resulting small sampling error. It is obvious that as the large errors, such as that of sampling, are diminished, more and more smaller disturbances will become noticeable, we expect this in our statistical and psychological material, just as we do in physical experimentation Thus, the second object of our work is that of furthering the investigation of error in tetrads, other than sampling error. As in physics, these errors should receive consideration in detail, before correlational material is submitted to complicated factor patterns of the kind given by Professor Kelley.[1]

---

* We express our deep indebtedness to Professor Spearman, under whom we, as Research Assistant, covered the present work

If, from our correlation tables for one thousand population, the tetrad-differences show error in excess of sampling error, then we seek explanations of the excess, and finally may need to make it a subject of special investigation. If the excess error has a reasonable explanation, if it makes contact with expectations, if it can be controlled by other experimentation, then the Theory of Two Factors, built upon the tetrad criterion, is still fully acceptable. Indeed, wherever and to the extent that the excess error can be shown to be reasonably expected, the theory would not be fully acceptable unless such error did occur Thus, the two objects set for the present paper—that concerning the *g*-factor for a battery of non-verbal subtests, for one thousand population, and that concerning error other than sampling error—really supplement each other.

A few sources of excess error in tetrads have been suggested already in previous papers. Excessive likeness and accidental linkages between abilities, and heterogeneity of race, age, and sex, were early recognized by Spearman as explanations of excess error. Some amongst verbal subtests (attributed to similarity of the fundaments or relations involved) was first observed by Davey,[2] and later by myself[10] (attributed to scholastic influences, in part when the population is drawn from different schools or districts). Professor Spearman[9] has noted the influence of different scales of scoring subtests, and, unless difficulties are obviated, estimates an error of 0.01 to 0.02 as likely to accrue for large populations if the subtests are variously scaled "Speed" preferences undoubtedly are a further potent source of excess error in tetrads. Idiosyncracies introduced into subtests when these are constructed by one psychologist may lead to error in the tetrads for these subtests,[10] and excess error may issue from a propinquity effect, deriving from the order of application of a battery of subtests.[10] Again, there is possible an influence entailed in testing by groups[7] Finally, not least of the influences that may lead to excess error in tetrads, there are mistakes in calculational work, test marking, and application. All the above are the kinds of possible disturbances in tetrads, influences leading to error in excess of sampling error Our object is to gather what information we can about these, and any, influences Some may be of psychological significance as objective specificalities, *i.e.*, as a group factor or factors, others may be trivial effects that further experiments could obviate.

EXPERIMENTAL MATERIAL

The data to be examined are for a non-verbal group test of eight subtests, applied by myself to 1037 girls. Testing was by groups, of about fifty girls per group, in elementary schools * Eleven schools, half the number in the region, were drawn from the centre of the city of Newcastle-on-Tyne (England). The schools, classes, numbers of girls per class, are shown in Table I The average ages of the children in the classes is given at the foot of the Table.

There were no extreme regional differences in social or economic status of the population of girls. The schools were much alike in point of teaching conditions and attainments. The small area covered, comprising the centre and greater part of the city, and the localizing effect of the city's administration, combined to give the population a distinct homogeneity in respect to influences of the kind just mentioned

Age distribution is given in Table II Eight hundred eighty-seven girls were drawn from Standards V and VIb (roughly equal classes), all of age falling within the range 10 years to 12½ years at the time of testing. These give to our work a large measure of homogeneity, so far as school can be compared with school, in both age and educational attainments or influences The various classes from which these girls came contained forty girls below ten years of age, and these were retained. To offset these young bright girls, we took fifty bright girls from Standards IVa, of age less than ten years, and a further seventy-six (at random from the same IVa classes) of age 10 to 12½ years Thus, one hundred twenty-six girls were from Standards IVa, but a large part of these were of prescribed age range (10 to 12½ years). Of the girls over 12½ years of age, seventy-five were in the various Standards V and VIb The girls entered from Standard VIIb of one school could be omitted without altering any of the results reported in the course of our work

---

* The author is indebted to the Committee and to Mr Walling, Director of Education.

TABLE I

School	Number of girls in				Total per school	Number of groups tested
	Standard IVa	Standard Va and b	Standard VIb	Standard VIIb		
A	28	40	22		90	2
B		51	40		91	2
C		69	31		100	2
D		58			58	1
E	33	65			98	2
F	19	58	41		118	2
G	14	36	39	23	112	2
H		50	47		97	2
I		34	53	1	88	2
J	6	61	27		94	2
K	26	65			91	2
Totals per standard	126	587	300	24		
Total all standards					1037	
Average age per girl	10 yr 2 mo	10 yr 10 mo	12 yr 1 mo	12 yr 5 mo		

TABLE II—AGE DISTRIBUTION FOR 1037 GIRLS

AGE (YEARS, MONTHS)	FREQUENCY
8- 6 to 8-11	5
9- 0 to 9- 5	21
9- 6 to 9-11	64
10- 0 to 10- 5	141
10- 6 to 10-11	156
11- 0 to 11- 5	206
11- 6 to 11-11	185
12- 0 to 12- 5	169
12- 6 to 12-11	60
13- 0 to 13- 5	24
13- 6 to 13-11	6

The non-verbal group test was applied in June 1929 Each testing group received the verbal group test on one day, followed by the non-verbal (our particular concern in the present paper) on the next day. Altogether, 1037 girls received both group tests.

The Non-verbal Subtests —The group test had to be easily applicable in forty minutes full testing time. The subtests therefore had to be of simple construction, and we anticipated rather low intercorrelations for some of the subtests; on the other hand we expected that two of the subtests would be fairly highly *g*-saturated (No. III and No V). The names of the subtests, in order of application to the girls, with the time allowance and number of test-units per subtest, were as follows:

I *Alphabet Construction* 3 minutes, 10 test-units
II *Code* 2 minutes, 8 lines of code, scored per half-line
III. *Fitting Shapes* 5 minutes, 15 test-units
IV *Picture Completion*· 3 minutes, 15 test-units
V. *Analogies, Form*· 3½ minutes, 16 test-units
VI. *Counting Cubes* 2½ minutes, 15 test-units.
VII *XO Series Completion.* 3 minutes, 15 test-units
VIII. *Overlapping Shapes* 3½ minutes, 24 test-units.

Throughout our work, we shall refer to these subtests by the Roman numerals I, II, etc. A brief description of each subtest follows:

In subtest I the testee was given a capital letter, and in imagination a piece had to be cut away, leaving as remainder any other capital letter If R was given, P could be the required response I's were debarred, the given letter could be unusually orientated (as A thought of as V), and one, two, or more pieces could be cut away (in imagination) The given letters were of simple printing, three-fourths inch high. A sample test-unit is shown in Fig 1, in which three different responses are expected The capitals used were, in order, X, U, A, H, E, S, Q, P, W, B, and the number of responses expected in each were 1, 1, 1, 2, 3, 2, 3, 5, 3, and 8, respectively In marking we allowed one mark each for test-units 1, 2, 3, and 6, and two marks for the rest, Nos 4, 5, 7, 8, 9, and 10 In the latter test-units one mark was allowed for half or more correct responses. In explaining the subtest, which appears to be readily understood, use was made of sample capitals Y, D, and M, providing V, L and C, and V and N, respectively The samples were worked through on a black-board, following a standard procedure

Subtest II was the Code of the National Intelligence Group Test, a sample of a similar kind of subtest being worked through on the black-board

Subtest III can be understood by referring to Fig 2 Three small shapes are given (together, to the left), which, when properly fitted together, form the larger shape (on the right) Lines had to be drawn in the larger shape (right) to indicate how it could be cut to give the three small shapes A correct response required *the drawing of no more than two or three lines in any test-unit* Four sample test-units were worked through on the black-board

Subtest IV was of the usual picture completion type, for example, a cup may be given which is minus its handle, and the testee is required to indicate the missing part

Subtest V, an Analogies subtest, made use of shapes and lines in various relations one to another A sample test-unit is given in Fig 3 The task of explaining the subtest's requirements was rendered easy by referring the girls to a verbal analogies subtest given to them on the previous day

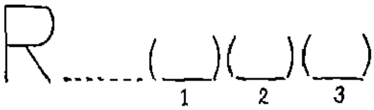


Figure 1

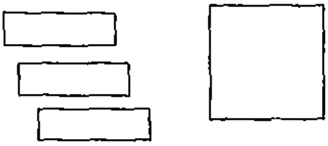


Figure 2.

Subtest VI was a modified form of the American Army subtest of this name

Subtest VII, a series completion subtest, can be understood by referring to the sample below

X X O X X O X X O X X O — — — —

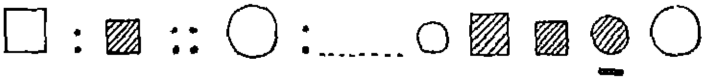


Figure 3

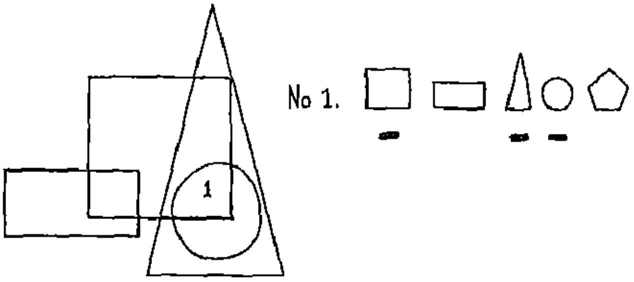


Figure 4.

The four dashes had to be filled in so as to continue the series, "X X O X" being the required responses in this sample.

Subtest VIII required a little more preliminary explanation than was needed for most other subtests. Figure 4 shows the type of test-unit. All the shapes overlapping the 1 had to be indicated by cancelling the appropriate shapes adjacent to No. 1 on the right A large square, a long rectangle, a long thin triangle, a circle, and a "five-sided" figure (regular pentagon), were the shapes made use of in each test-unit. In the case of the sample shown in Fig. 4, the square, triangle, and circle, required cancelling.

The subtests have been described for adult understanding, but the testing directions were much more childlike and concrete Application was standardised to an easy facility, and the subtests were given with this regularity throughout. Each subtest was prefaced by a fore-practice of six or so test-units, worked through on the class black-board, altogether fifteen minutes of the forty minutes testing time was so spent in giving directions and working the sample test-units. For each subtest the girls were allowed about fifteen seconds to settle to be ready for the command "Go," each girl fingering the page ready for turning (cyclostyled sheets, stapled in the top left-hand corner, are quickly turned over, leaving the subtest exposed), and were brought to a stop with an energetic "Stop" command I had group-tested many hundreds of boys and girls previously.

Subtests I, II, IV, and VII, occupied one sheet of paper each (quarto size), while the rest required two pages each; but in the latter cases the two pages of test-units faced each other, and appeared to the girls as one double-sized page in each case. There was, therefore, no page-turning during the time allotted for working these subtests.

Enquiries elicited that no group tests had been applied to these girls previously.

INTERCORRELATIONS AND TETRAD-DIFFERENCES

Intercorrelations for the eight non-verbal subtests were first calculated for crude scores, using the formula for differences (Kelley[6] p 180) All calculations were checked in various ways, but the question of mistakes receives attention later. Table III gives the product-moment correlations for the eight non-verbal subtests, with age, for 1037 girls. These we are to consider in terms of the Spearman Theory of Two Additive Factors.

When we are concerned with tetrad-differences for a table of intercorrelations, we calculate one-half the number of tetrad-differences possible for the table (the other half are identical in value, but of opposite sign) The *mean* of this half-number of tetrad-differences is thus the average or mean deviation about the mean of the observed differences.

For the theoretical PE of the tetrad-differences we use formula 16A (Spearman,[7] p. xi) For exact comparison with this theoretical PE we calculate the observed probable error, given by $0.6745\sqrt{\Sigma t^2/n}$, where t stands for "tetrad-difference," and n is the number of tetrad-differences.

If "normal" probability distribution obtains for the observed tetrad-differences, the above *mean* deviation about the mean of the observed tetrad-differences, multiplied by 0 8453, gives the value of the probable error for these observed tetrad-differences We anticipate that this latter value should be approximately the same as that given by the sigma above

For any set of tetrad-differences, then, we can give (and it is our usual practice to do so throughout our work) the following values·

(*a*) The *mean* of the half-number of tetrad-differences, sign being disregarded This is the mean deviation about the mean, or average, of the full set of tetrad-differences, when regard is paid to sign, the average of the full set being zero

(*b*) The probable error of the differences, calculated from the above *mean*, i e, from the mean deviation, i.e,

$$pe = 0\ 8453 \times \mathit{Mean}\ \text{deviation}$$

(*c*) The probable error of the differences, calculated from the observed sigma of the differences, i e,

$$pe = 0\ 6745\sqrt{\frac{\Sigma t^2}{n}}$$

(*d*) The theoretical probable error, always expressed as PE, given by equation 16A (Spearman).

From Table III, before age is partialled-out, we obtain the following results for the tetrad-differences.

(*a*) *Mean* of 210 tetrad-differences .	0 0232
(*b*) Observed pe, from *mean* × 0 8453 .	0 0196
(*d*) Theoretical PE	0 0095

If we partial out the age correlations, the new intercorrelations for the eight non-verbal subtests give the following results

(*a*) *Mean* of 210 tetrad-differences, age partialled-out correlations	0 02265
(*b*) Observed probable error, from 0.8453 × the above *mean*	0.0191
(*c*) Observed probable error, from 0 6745 × sigma	0 0184
(*d*) Theoretical PE	0 0094

Thus, as a whole, the intercorrelations show error in excess of that expected as sampling error, the excess being of the order 0.016 Further, age has no significant influence on the tetrads here considered. Our immediate problem, then, is the attempt at location of this residual excess error

LOCATION OF THE EXCESS ERROR

When the tetrad-differences show excess error it is our first object to try to explain the excess, in terms of specificality, *i.e.*, of a group factor or factors, or in terms of extraneous influences, such as calculation mistakes. Some possible specificalities and influences have been described in the Introduction

Without going into details, it may be said here that there is no evidence that any of the influences mentioned above (with exceptions to be considered below) have singly produced the residual error. Thus, the effect of testing by groups was found by me to be here without influence on the tetrads. Again, subtests III, V, VI, and VIII have somewhat similar shapes as fundaments, but there is no evidence of a broad gross specificality because of these fundaments Subtests III and VIII involve very similar geometrical figures as fundaments, but specificality is not shown by $r_{\mathrm{III,VIII}}$ We guarded against "propinquity" influences in our testing procedure, although we would not be quite free from an habituation disturbance for $r_{\mathrm{I,II}}$ or $r_{\mathrm{I,III}}$ From the way in which the interesting subtests (judged from the girls' attitudes towards the tests) were introduced in the battery, and from the short testing time entailed, we can neglect possibilities of fatigue, subjective or otherwise, entering our material No matter what we attempt, we can not isolate any single correlation and submit that it is associated with the larger, or greater part, of the excess error shown by the tetrads. Thus, in a preliminary search for possible disturbances there remained two for consideration, first, the possible calculation mistakes, second, the possible error introduced by dissimilar scales, by faults in score distributions.

If, by recalculating intercorrelations for new scales (a new distribution of scores for each subtest, without distorting the sense of the scores), we finally rid the tetrad-differences of excess error, the whole procedure will be open to the criticism that the intercorrelation tables vary slightly one from another, due to calculation mistakes, so that the final removal of the excess error would be considered to be merely a happy chance in the calculations. But, at least, we shall see the extent within which such variations due to calculation mistakes can be taken to be of influence in our work; and, should our procedure show orderly diminution of the observed tetrad-differences, some evidence will have been obtained for a fair measure of soundness in the calculations, and for the supposed influence of the irregular distributions of scores

The Subtests Rescaled.—We proceeded in two stages. First it was thought perhaps sufficient to rescale the subtests so as to give approximate "normal" distributions of scores for the 1037 girls When it was found that this approximation failed to give the expected improvement in the error shown by tetrads, we subsequently converted all crude scores into a "standard" normal distribution, the same for each subtest. Using new correlations for seven of our subtests, each subtest having an approximate "normal" score distribution, gave a *mean* of 105 tetrad-differences of amount 0.0229; this is to be compared with the value 0.0234 observed for the same seven subtests with the intercorrelations of Table III The average of the twenty-one intercorrelations for the new table was 0.3598; the corresponding average for Table III is 0.3614 (subtest VI omitted). Thus the rescaling to an approximate "normal" distribution (amounting to removal of any "skewness"), does not rid us of the excess error in the tetrads. Nevertheless it perceptibly altered the intercorrelations, in the direction of a smoothing-up of the table of intercorrelations, giving better "hierarchy." It appeared that just another such slight smoothing would bring the tetrad-differences to within sampling error values

TABLE III.—PRODUCT-MOMENT CORRELATIONS FOR N OF 1037. CRUDE SCORES

	Age	I	II	III	IV	V	VI	VII	VIII
Age		1783	1066	0550	0866	0039	0443	0152	0813
I			3565	4181	3227	3816	2254	2004	3376
II				3275	3155	3008	1609	3083	3802
III					3998	4759	3663	3473	3678
IV						3810	2671	2424	3152
V							2777	4045	3098
VI								2618	2110
VII									3251
VIII									

It is not expected that crude scores should have much more exact "normal" distributions of scores than those obtaining for our subtests. But the number of girls given zero score in particular subtests is frequently disproportionately large and, when the test-units of a subtest are not sufficiently smoothly graded, the short time-allowance per subtest tends to accentuate the frequency of certain scores. Disturbances of this kind may hide the general results that are our concern. To take into consideration influences of this kind we considered it worth while trying our data with new scales, this time fitting each subtest to a more exact "normal" distribution, the same for each subtest.

For convenience the six subtests I, II, III, IV, V, and VIII were made the subject of the new conversion. The crude scores which supplied Table III were now converted to fit exactly the following approximate "normal" distribution.

Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Frequency . .	3	7	16	24	48	70	98	118	135	134	118	98	70	48	24	16	7	3

The method used in converting the crude scores will be understood by referring to the following example of the procedure.

The crude scores and frequencies for subtest I were as follows:

Crude scores	0	1	2	3	etc.
Frequency . .	2	10	27	32	etc.

Of the ten girls obtaining crude score 1, one was chosen at random, and her score was reduced to zero [making, with the two having zero crude score, the three new-scale 0 scores required for the "normal" distribution (for convenience we name this distribution the "standard" one)] seven of the girls retained their score 1 (now taken as new-scale), and the other two received new-scale score 2. We now require fourteen more new-scale 2's. these were taken at random from amongst the twenty-seven having crude score 2, leaving thirteen. The thirteen are accredited with a score 3 on the new-scale. Similarly eleven of the thirty-two crude score 3's are given new-scale 3's, and the rest are given new-scale 4's. This process is repeated up to the maximum score, 17 new-scale, all the selections within a particular crude-score being made at random. It is obvious that the new scaling is reasonable in that it does not distort the sense of the subtests, and it is valid, having in mind the arbitrary system of allotting crude scores

The six subtests so rescaled provide the intercorrelations given in Table V, age being neglected. (It seems that age influence, for all our tables of correlations, would introduce at most about 0.005 error in the tetrads. See in this connection Holzinger,[4] p. 28.) The subtests give tetrad values as follows:

(*a*) *Mean* of 45 tetrad-differences for Table V	0.0188	(β)
(*b*) Observed probable error, given by *mean* × 0.8453	0.0159	
(*c*) Observed *pe*, given by sigma × 0.6745	0.0171	
(*d*) PE	0.0096	

These values show improvement when compared with corresponding tetrads for Table III. But the improvement is greater than the above observed probable error evinces. One correlation, $r_{\text{II,VIII}}$, has associated with it the larger of the tetrad-differences. If we omit the twelve tetrad-differences involving this correlation, we are left with the following:

(*a*) *Mean* of 33 tetrad-differences for Table V, omitting $r_{\text{II,VIII}}$	0.0104	(α)
(*b*) Observed probable error, given by *mean* × 0.8453	0.0088	
(*c*) Observed *pe* given by sigma × 0.6745	0.0080	
(*d*) PE	0.0095	

The observed probable error (0.0088) is now attributable solely to sampling error. It seems that the improved distribution of subtest scores has removed error in the tetrads, other than sampling, and that due to $r_{\text{II,VIII}}$, as we see from a comparison of the values at (α) with corresponding tetrads for Table III. Thus the forty-five comparable tetrads for Table III (crude scores) have the following values:

(*a*) *Mean*, forty-five tetrad-differences for Table III	0.0244
(*b*) Observed *pe* (*mean* × 0.8453)	0.0206
(*d*) PE	0.0100

Again, if we omit from these forty-five tetrads the twelve that involve the correlation $r_{\text{II,VIII}}$, we are left with the following values, for comparison with the observed probable error of amount 0.0080 for Table V:

(*a*) *Mean* of thirty-three tetrad-differences, for Table III, subtests I, II, III, IV, V, and VIII, omitting $r_{\text{II,VIII}}$	0.0191
(*b*) Observed probable error, given by *mean* × 0.8453	0.0159
(*c*) Observed probable error, given by sigma × 0.6745	0.0169
(*d*) PE nearly	0.0098

Values quite similar to these for Table III were obtained also for the table of intercorrelations alluded to at the beginning of this section, where seven subtests were rescaled to an approximate "normal" distribution of scores. The value 0.0169 at (c) above is for crude score correlations, whilst 0.0080 at (α (c)) is for "standard" normal probability distributions of subtest scores; we conclude that the improved distribution of scores has removed excess error in the tetrads (excepting $r_{\text{II,VIII}}$). The influence of the crude scores is likely to be of amount 0.0149 (given by $\sqrt{0.0169^2 - 0.0080^2}$). This, we note, is of amount estimated by Professor Spearman as likely to accrue for large populations if subtests are variously scaled as mentioned in the Introduction to this article.

TABLE IV.—MEANS AND SIGMAS OF THE NON-VERBAL SUBTESTS, FOR N = 1037, GIRLS

Subtest	Mean	Sigma	Maximum score
I	6.89	2.32	15
II	7.01	2.22	15
III	4.48	1.87	15
IV	4.64	1.88	10
V	7.07	3.00	16
VI	3.81	2.47	13
VII	7.33	1.90	13
VIII	8.07	2.69	17

We conclude, then, that the observed probable error given at (α) above for Table V, is as low as we can reasonably work for, and the tetrads provide sampling error differences only; to complete this conclusion, however, we must explain the excess error apparently shown by the correlation $r_{\text{II,VIII}}$. Further, it appears that the rescaling, to fit a "standard" "normal" distribution of scores, has effected the improvement in the error shown by the tetrads. In this connection we should note the following contributory evidence. The means of the corresponding intercorrelations in Tables III, that alluded to for seven subtests with approximate "normal" distributions, and V, are 0.3614, 0.3598, and 0.3527 respectively—these values, we suggest, offer an example of the fair quality of the calculations, and this receives confirmation in other directions, for instance, from the objective specificality (considered below) for $r_{\text{II,VIII}}$. We can not readily consider that a happy chance has resulted in the improvement shown in the observed error at (α). Calculation errors there are, no doubt, but they are too fine to be significant in our tetrads

To complete our account of the factor content of the non-verbal subtests we require to explain the excess error apparently shown by $r_{\mathrm{II,VIII}}$.

Consideration of the $r_{\mathrm{II,VIII}}$ Specificality.—One would not place too much regard upon a single case of divergence from sampling error values, and usually the details we are about to give would scarcely be necessary. But, the excess error attributed to $r_{\mathrm{II,VIII}}$ can receive a simple explanation, and the specificality can be shown to be objective, it is a matter of importance to report the data concerned, for reasons connected with the acceptability of the Two Factor Theory for our data, for the contact it makes with the investigation of error other than sampling error, and for the evidence provided of the objectivity of the correlational values under consideration.

TABLE V—PRODUCT-MOMENT CORRELATIONS NON-VERBAL SUBTESTS, WITH SCORES CONVERTED TO A "STANDARD" NORMAL PROBABILITY DISTRIBUTION
N = 1037 GIRLS

Subtest	II	III	IV	V	VIII
I	3463	4121	3274	3733	3207
II		3484	2074	3121	3707
III			4110	4711	3508
IV				3771	3012
V					3278
VIII					

We have suggested that $r_{\mathrm{II\,VIII}}$ is the source of excess error in Table V We do so by noting that the largest tetrad-differences for Table V uniformly involve this correlation Tetrads of the following type are then isolated, and omitted from the main body of tetrads

$$\left.\begin{aligned}(r_{AB}\ r_{xy}) - (r_{Ax}\ r_{By}) &= f\\(r_{AB}\ r_{xy}) - (r_{Ay}\ r_{Bx}) &= f'\end{aligned}\right\} \qquad (1)$$

(where r_{AB} is suspected for excess error, and where x and y are any other subtests, taken two at a time, regard being taken of the sign of the difference). The twelve tetrads of this form for $r_{AB} = r_{\mathrm{II,VIII}}$, for Table V, have mean of amount +0.0420.

Now, there is evidence from other sources that strengthens the claim of the $r_{\mathrm{II\,VIII}}$ excess error to a fair degree of objectivity The evidence is as follows:

In the table for crude scores, Table III, $r_{\mathrm{II,VIII}}$ has the largest mean associated with it, for tetrads of form (1), namely an amount +0.035. Again, correlations worked for a sub-population of two hundred girls, a selected group with ages ranging 10-9 years to 11-3 years at the time of experiment, gave the following results:

(*a*) *Mean* of two hundred ten tetrad-differences, sub-population of two hundred, eight non-verbal subtests	0.0319
(*b*) Observed probable error, given by *mean* × 0.8453	0.0270
(*d*) PE	0.0230

There are thirty tetrads of form (1) for $r_{\mathrm{II,VIII}}$ in this two hundred ten tetrads, with mean of value +0.0737; the omission of these thirty from the full table of tetrad-differences leaves the following results:

(*a*) *Mean* of one hundred eighty tetrad-differences, sub-population of two hundred girls	0.0248
(*b*) *Mean* × 0.8453	0.0210
(*d*) PE	0.0230

Other correlation tables, for sub-populations of one hundred girls, and five hundred girls, gave similar results. The most noticeably large tetrad-differences always involved the correlation $r_{\mathrm{II,VIII}}$, and the omission of this correlation left tetrads with sampling error values only. The result shown for Table V, then, is not particular to that table of correlations. The excess error associated with $r_{\mathrm{II,VIII}}$ is independent of the scoring methods used in our work; and the observation of its influence is undoubtedly repeatable, *i.e.*, objective.

An acceptable specificality between the two subtests II and VIII can be explained in terms of a "quantity-work," or "speed," preference. The Code subtest (II) may be liable to a pronounced preferential attack for "speed" as against "quality." Subtest VIII would seem to entail a similar preference on the part of the testees. One girl may diligently ensure that each test-unit in subtest VIII is completely and surely answered, *all* the overlapping shapes being sought for. Another girl may be satisfied if one or two overlapping shapes are found, and would spend no time in a search for further shapes. The former may respond to five test-units only, marking probably fifteen figures that overlap, without a single error or omission; the latter may respond to all twenty-four test-units, marking forty overlapping figures, but with not a single test-unit correctly answered. A differential influence of this kind has been observed frequently in other work, in researches made in our laboratory. With our present material, we have a test of the "explanation", and we next give this attention.

Subtest VIII was scored for quality (Q, completeness of answer), and for quantity (q, total number of shapes correctly indicated, corrected for chance correct responses). The scores were added, giving equal weight to each, and $(Q + q)$ was then the crude score for subtest VIII, used in all work described in the previous sections. The separate Q and q scores, however, give us the following additional data:

(1) The correlations $r_{II,q}$ and $r_{II,Q}$ are 0.4088 and 0.3217 respectively.

(2) r_{Qq} is 0.8505.

(3) Using the Q score only, the following new correlations were calculated, all scores being set on the "standard" normal distribution excepting Q

$$r_{Q,I},\ r_{Q,II},\ r_{Q,III},\ r_{Q,IV},\ r_{Q,V},\text{ having values}$$

0.3230, 0.3217, 0.3216, 0.2652, and 0.3232, respectively. Replacing the correlations for subtest VIII in Table V by the corresponding ones given above, gives a new set of tetrad-differences, with value as follows.

(*a*) *Mean* of 45 tetrad-differences, Table V, having Q instead of VIII	0.0152	(γ)
(*b*) Observed probable error, given by *mean* × 0.8453	0.0128	
(*c*) Observed probable error, given by sigma, × 0.6745	0.0128	
(*d*) PE nearly . .	0.0095	

The twelve tetrads of form (1) for the correlation $r_{II,Q}$ have mean of value 0.0251. The above values should be compared with the values given at (β) on page 178; and the mean 0.0251 likewise should be compared with the mean of 0.0420 given on page 180, for the tetrads of form (1) involving $r_{II,Q}$ and $r_{II,VIII}$ respectively. The use of Q gives lower error in the tetrads.

(4) With the correlations already provided, it is possible to calculate the following correlations

$$r_{I,q},\ r_{II,q},\ r_{III,q},\ r_{IV,q},\ r_{V,q}$$

These have values 0.2940, 0.4088, 0.3519, 0.3142, 0.3074, respectively. As at (3) given, we replace the correlations for the subtest VIII in Table V by these q correlations. The resulting table of correlations provides forty-five tetrad-differences, with value much the same as those given above at (3) for the Q correlations, except that the twelve tetrads of form (1) have mean +0.0545. This value should be compared with the 0.0251 for the Q-measure tetrads of form (1), and with the $(Q + q)$ measure value of 0.0420.

Thus, it seems, from the considerations we have given of the matter in the above paragraphs, that the specificality shown by $r_{II,VIII}$ has a fair degree of satisfactoriness, owing most to a common influence between subtest II and the q (quantity) measure of the subtest VIII. The excess error for the tetrad-differences of Table V, which we associated with $r_{II,VIII}$, is shown above to be repeatable and therefore objective (within the limits of our data for the 1037 girls); similar error has been found in other work, in which a similar explanation in terms of a "speed" effect has been offered; the explanation seems reasonable for the two subtests concerned; and it receives experimental support from the Q and q correlations. Thus, the $r_{II,VIII}$ specificality is well authenticated.

Further, this "speed" specificality seems to be the *only* one that can be isolated for Table V. Relevant to this conclusion we should add the following details.

One may envisage a complicated cancelling, in the tetrads for Table V, whereby the correlations $r_{I,III}$, $r_{III,IV}$, $r_{III,V}$, and $r_{II,VIII}$, cancel out (except for the latter) any influence that each may have separately towards excess error. If we omit all tetrads involving these four correlations, we are left with only eleven of the forty-five tetrads for Table V, with *mean* of amount 0.0083, observed probable error of 0.0069, theoretical PE of 0.0094. But we are not justified in suggesting the isolation of correlations that already give only sampling error value to tetrads. However, so far we have been critically concerned with only six of our eight non-verbal subtests; if need be, we could widen our data by including the two subtests so far not *rescaled, thereby perhaps obtaining further information about the* effect of any, or all, of the four correlations considered above. There is reason to suppose that if subtests VI and VII were rescaled, they, too, would fall into line with the values and characteristics shown by Table V; thus, subtest VI was included in the correlation table alluded to in this article under "The Subtests Rescaled" (for the seven subtests with approximate "normal" distribution of scores), and it fits the tetrads as well as any other subtests; and, further, subtests VI and VII had clearly the most irregular crude score distributions, which amply explain the correlation anomalies they show in Table III. Until we find clear need to seek further information about the abovementioned correlations, or any others, we propose being satisfied with the data obtained already for the six subtests in Table V.

The correlation $r_{III,V}$ requires a moment's consideration because of contact made with it in the course of work to be described in a subsequent paper. We may state, without going into details (a matter that can be examined by anyone so interested, since all the necessary data are provided in this paper), that $r_{III,V}$ is in no way singular in its effects in tetrads. No matter from what angle we examine our data, we obtain no evidence in support of specificality for $r_{III,V}$. Parenthetically we should add that in any detailed examination it should be remembered that the Q measure for subtest VIII is biased for *quality* (with the narrow conception given to this word, namely, "completeness of answer"), and this slightly disturbs correlations with Q Any need to consider $r_{III,V}$ (or $r_{I,III}$, $r_{III,IV}$ etc.) as a possible source of excess error, is removed if the Q or q scores are neglected

We conclude that $r_{II,VIII}$, in Table V, alone involves acceptable specificality.

Résumé and Conclusion

Commencing with the intercorrelations of Table III, which provide an observed probable error of 0 0196 for their tetrad-differences, we have endeavored to locate error of amount about 0.016 in excess of that expected for sampling error (0.0095). If we were able to explain the excess error satisfactorily, then we are left to conclude that the intercorrelations for the non-verbal subtests have good agreement with the Theory of Two Additive Factors

Of first importance is the question of our calculation mistakes. The sequence of error for tetrads for Table III, that alluded to for the seven subtests with only approximate "normal" score distributions, and for Table V, as well as the play of error for the various influences considered in the previous section are the evidence that we offer of the quality of our correlational calculations. We cannot hope to be free entirely of calculational error; but, we add, the correlation calculations should be accepted even if our tetrad calculations are found slightly inaccurate, because the correlations have received the greater attention Further, our data are lodged permanently in the Psychological Laboratory, University College, London, and are available for anyone to rework.

Neglecting possible calculation mistakes, and age influence (which is no doubt present, but introduces very little error, of amount 0 005 at most), we account for the excess error of Table III tetrads as follows.

First, there is error introduced when the subtest scores are not distributed satisfactorily "normally." When the crude scores are converted to fit a "standard" normal probability distribution, the excess error is removed, except for a single specificality that can be isolated and explained. Neglecting this single specificality gives an observed probable error of 0.0081, for a theoretical PE of 0.0095.

We show that this single specificality, for $r_{II,VIII}$, is objective, and acceptable as the only specificality among our non-verbal subtests.

The Theory of Two Additive Factors therefore fits well with the observed results for our non-verbal subtests, allowing for $r_{II,VIII}$. We explain the intercorrelations in terms of a common g-factor, and factors specific to each subtest (with *specificality* for $r_{II,VIII}$). The g-factor we take to be that first observed by Professor Spearman, the factor that appears to characterize relation and correlate eduction. Thus, the g-factor hitherto observed for non-verbal subtests for small populations, is here shown to obtain for one thousand population.

Finally, concerning our second object, that of knowledge of error other than sampling error, we can here say that the facts can best be given on completion of our work, with the verbal subtests considered in addition to the non-verbal. At this juncture we note that, accepting our correlations, it is possible to observe the influence of non-normal distribution of scores for the subtests. Our second point of note is that we isolate a single specificality, attributed to a "speed preference." Such a "speed" effect must be looked for in all work with g-tests; and the influence must be kept in mind particularly in our work with the verbal subtests.

REFERENCES

1. Kelley, T. L. "Crossroads in the Mind of Man," 1927.
2. Davey, C. M. A Comparison of Group Verbal and Pictorial Tests of Intelligence. *British Journal of Psychology*, Vol. XVII, 1926, pp. 27–48.
3. Fortes, M. Thesis, Library of University of London.
4. Holzinger, K. J. "Statistical Résumé of the Spearman Two-Factor Theory." University of Chicago Press, 1930.
5. Kelley, T. L. "Statistical Method." The Macmillan Company, 1924.
6. Line, W. Thesis, Library of University of London.
7. Spearman, C. "Abilities of Man." The Macmillan Company, 1927.
8. Spearman, C. Factor School of Psychology, Part X. *Psychologies of* 1930, pp. 339–366.
9. Spearman, C. Disturbers of Tetrad Differences. *Journal of Educational Psychology*, Vol. XXI, No. 8, 1930, p. 559.
10. Stephenson, W. Thesis, Library of University of London.

A COMPARISON OF WHITE AND NEGRO CHILDREN: NORMS ON MIRROR-DRAWING FOR NEGRO CHILDREN BY AGE AND SEX

R. J. CLINTON

Oregon State College

A great deal has been written and said by way of comparing the white and Negro races, both in mental ability and motor ability. Ragsdale[1] said that there was not much difference in the ability of white and Negro boys, in the rate of tapping, "although there seems to be a slight tendency for white boys to be superior at the later ages." He found that there was not much difference in the rate of tapping of white and Negro girls until the ninth year when "white girls are consistently superior in tapping rate." Waltner[2] states: "This study of the learning capacity of Negroes as compared with whites finds the power to form sensory-motor coordinations by Negroes to be about 72.8 per cent that of whites."

Bardin[3] observed that the anatomy and physiology of the Negro's brain hindered him in learning. Jordon's[4] personal observation was that the learning capacity of the Negro could be determined by the darkness of his skin. Phillips[5] suggested retardation as a criterion for judging the ability of Negro children.

Schwegler and Winn[6] studied a group of 58 Negro boys and girls and a like number of white boys and girls in the Junior High School of Lawrence, Kansas. The selection was by chance. They state:

> There seems to be an unmistakable difference in the intellectual life of the two groups studied. The median intellectual endowment of the colored group is about eight-five per cent of that of the white group.

---

[1] Ragsdale, C. E. "Psychology of the Negro." Thesis, University of Missouri, p. 144.

[2] Waltner, Erma. "Psychology of Negro." Thesis, University of Missouri, p. 36.

[3] Bardin, J. Factors in the Southern Race Problem. *P. Sc. M.*, 1913, Vol. LXXXIII, pp. 368–374.

[4] Jordon, H. E.: The Biological and Sociological Worth of the Mulatto. *P. Sc. M.*, Vol. LXXXIII, pp. 573–582.

[5] Phillips, B. A. Retention in Elementary School of Philosophy. *Psychological Clinic*, Vol. VI, p. 79.

[6] Schwegler, R. A., and Winn, Edith. Comparison Study of White and Colored Children. *Journal Educational Research*, Vol. II, pp. 838–847.

The writer became interested in the studies cited and desired to make a comparison of the abilities of white and Negro children[1] in mirrow-drawing, motor speed as shown by marking, and letter-forming speed as shown by writing, as well as to make some comparisons of the mental abilities of the two races

By means of the use of the Otis Self-administering Test of Mental Ability, the writer made a comparison of the mental abilities of an unselected group of white high school pupils, and an unselected group of Negro high school pupils. The one hundred fifty-five white high school pupils gave a mean IQ of 100 5, and the one hundred twenty-two Negro high school pupils gave a mean IQ of 84 5

METHOD OF THE STUDY

The mirror-drawing equipment[2] consisted of a stationary mirror on a drawing board, and a small shield to prevent the subject from looking directly at the hand while working The mirror-drawing pattern consisted of a two-circle pattern with numbers from 1 to 24. Numbers 2, 5, 8, 11, 14, 17, 20, 23 make up the small inner circle, and numbers 1, 3, 4, 6, 7, 9, 10, 12, 13, 15, 16, 18, 19, 21, 22, and 24

TABLE I.—COMPARISON OF WHITE BOYS AND NEGRO BOYS FROM SIX TO SEVENTEEN YEARS OF AGE IN MIRROR-DRAWING

White boys			Negro boys		
Age	Number	Mirror-drawing	Age	Number	Mirror-drawing
6	38	4 7	6	19	1 7
7	35	6 6	7	21	2 9
8	49	9 1	8	25	8 2
9	27	9 4	9	28	6 8
10	43	10 5	10	22	7 2
11	30	14 9	11	21	9 7
12	40	14 7	12	28	8 3
13	32	15 8	13	28	8 2
14	44	20 1	14	10	16 2
15	34	22 4	15	23	10 8
16	46	24 9	16	16	9 3
17	49	30 8	17	10	16 5

---

[1] The writer included 591 Negro children in this investigation

[2] Clinton, R J Nature of Mirror-drawing Ability Norms on Mirror-drawing for White Children by Age and Sex *Journal Educational Psychology*, Mar., 1930

make up the large outer circle. The subjects were instructed to draw a continuous line from 1 to 24, cutting the small dot at each number.

TABLE II.—COMPARISON OF WHITE GIRLS AND NEGRO GIRLS FROM SIX TO SEVENTEEN IN MIRROR-DRAWING

White girls			Negro girls		
Age	Number	Mirror-drawing	Age	Number	Mirror-drawing
6	28	2.9	6	17	1
7	41	4.9	7	18	4.9
8	53	5.9	8	24	3.7
9	28	8,3	9	32	4.6
10	44	8.9	10	37	6.9
11	49	10.9	11	20	5.3
12	40	13.8	12	28	7.2
13	37	16.1	13	45	8.1
14	57	22.6	14	18	7.7
15	61	30.7	15	30	9.2
16	48	34.4	16	24	23.1
17	49	38.6	17	15	20.1

TABLE III.—COMPARISON OF WHITE BOYS AND NEGRO BOYS FROM SIX TO SEVENTEEN IN MARKING SPEED AND LETTER MAKING SPEED

White boys				Negro boys			
Age	Number	Marks	Letters	Age	Number	Marks	Letters
6	38	119	54	6	10	98	45
7	35	124	74	7	21	113	71
8	19	153	123	8	25	153	84
9	27	160	150	9	28	169	118
10	43	195	179	10	22	167	141
11	30	203	191	11	21	192	146
12	40	223	200	12	28	216	160
13	32	243	228	13	28	222	183
14	44	252	234	14	10	229	197
15	34	264	241	15	23	254	209
16	46	271	250	16	16	263	234
17	49	285	268	17	10	248	230

The score was the number of lines completed in a five minute period. The writer designed a mirror-drawing pattern to use in Grades I, II, and III which included the figures from 1 to 15 The figures were so placed that the small children had no difficulty in locating them.

TABLE IV —COMPARISON OF WHITE GIRLS AND NEGRO GIRLS FROM SIX TO SEVENTEEN IN MARKING SPEED AND LETTER MAKING

White girls				Negro girls			
Age	Number	Marks	Letters	Age	Number	Marks	Letters
6	28	112	52	6	17	102	66
7	41	147	96	7	18	136	105
8	53	164	137	8	24	161	122
9	28	193	176	9	32	175	138
10	44	212	194	10	37	202	163
11	49	217	207	11	29	208	156
12	49	228	227	12	28	231	199
13	37	249	250	13	45	223	206
14	57	263	268	14	18	236	230
15	61	270	272	15	30	238	231
16	48	271	274	16	24	264	250
17	49	274	280	17	15	265	251

TABLE V —NORMS FOR NEGRO BOYS AND NEGRO GIRLS IN MIRROR-DRAWING, MARKING, AND LETTER MAKING

Negro boys				Negro girls			
Age	Mirror-drawing	Marks	Letters	Age	Mirror-drawing	Marks	Letters
6	1 7	98	45	6	1	102	66
7	2 9	113	71	7	4 9	136	105
8	8 2	153	84	8	3 7	161	122
9	6 8	169	118	9	4 6	175	138
10	7 2	167	141	10	6 9	202	163
11	9 7	192	146	11	5 3	208	156
12	8 3	216	160	12	7 2	231	199
13	8 2	222	183	13	8 1	223	206
14	16 2	229	197	14	7 7	236	230
15	10 8	254	209	15	9 2	238	231
16	9 3	263	234	16	23 1	264	250
17	16 5	248	263	17	26 1	265	251

The two mirror-drawing patterns were evaluated on comparable groups in order to get continuous norms through all ages from six to seventeen

The Tables I and III show a comparison of the ability of white and Negro boys from six to seventeen years of age in mirrow-drawing, and in making speed and letter making. Table II and IV show a comparison of the same abilities between white and Negro girls from six to seventeen years of age Table V shows the norms in mirror-drawing, marking speed, and speed in making letters for Negro boys and girls from six to seventeen

CONCLUSIONS FROM THE STUDY

1. Unselected white high school pupils are superior mentally to unselected Negro high school pupils, as shown by the tests

2. In the simple motor process of marking, there is not much difference between the white and Negro children.

3. In writing, which requires a greater degree of motor-sensory coordination, the superiority of the white children is clearly shown.

4. In the complex motor-sensory coordination process of mirror-drawing, the white children are consistently superior to the Negro children.

5. The ability to do mirror-drawing increases rather consistently from year to year with Negro children, as well as with white children.

A STUDY IN REVERSING THE HANDEDNESS OF SOME LEFT-HANDED WRITERS[1]

NORMA V SCHEIDEMANN

University of Southern California

AND

HAZEL COLYER

Los Angeles City Schools

The subject of handedness has long been of interest to psychologists, neurologists, eugenicists, and popular writers. Reference to professional literature will show that the classroom teacher of the primary grades, who has the best opportunity to observe hand preferences in non-established or poorly established activities, has taken no especial interest in the subject. The majority of elementary teachers of a decade ago seemed to have a single rule in establishing writing handedness—all children must write with the right hand. Many left-handed adults tell us of excruciating ordeals they were required to undergo in trying to "break up" left-handed tendencies The majority of elementary teachers of today likewise seem to have a single rule in regard to establishing handedness in writing—the right-handed child should write with his right hand and the left-handed child should be permitted to write with his left hand.

That a child upon entering school, in his eagerness to do exactly the right thing, may be influenced by unimportant factors so that he may show an unnatural hand preference for writing movements is usually not recognized. A child's lack of critical judgment is manifested constantly in his imitation of non-consequential factors in hopes of attaining significant traits or habits. Thus, a child may quite confidently try to smoke in order to become manly like father, or try to walk pigeon-toed in order to be like a greatly admired friend Imitation of hand preference for writing may also be made by con-

---

[1] This study was made during the second semester of 1929–1930 in the second grade (IIB) of the Melrose Avenue School, Los Angeles City Schools. This school is located in a residential district of Hollywood The children are normal children and come from much above average homes

Mrs Colyer, the second grade teacher of the group studied, interviewed the parents of the children and carried out all the remedial procedure in establishing a reversal of handedness in writing scious imitation of an admired playmate or teacher or by unconscious imitation when the child is anxiously trying to follow, in most minute detail, every movement of one demonstrating a desired habit or skill. That first grade children may show unnatural hand preferences for writing and that the primary teacher may unwittingly mistake these unnatural preferences for innate dominance, was revealed in a recent study of a group of left-handed writers [1]

Of a second grade group of thirty-four children, sixteen were left-handed writers Both the first and the second grade teachers of the group were left-handed; the children's writing habits were established before they entered the second grade. The improbability of so high a percentage of left-handed children in an unselected group of normal individuals led to a study of these children in order to determine, if possible, the native handedness of each left-handed writer and to effect a reversal of handedness in cases where that might seem warrantable.

Since no single test has been devised whereby a child's native handedness can be definitely discovered, each child was given a series of tests, one for eyedness and nine for handedness Native eyedness is perhaps the best single indication of a child's natively dominant side. Study of bodily asymmetry has resulted in the recognition of dominance of one side of the body over the other, that is, individuals are either dextroexpert, generally, as to ear, eye, hand, and foot, or else they are sinistroexpert [2]

Hand- and foot-preferences are subject to training, but eye preference, except for accident or disease, remains uninfluenced by experience. Hence the eye test was considered the criterional test for native handedness, but it was not considered justifiable to effect a reversal of handedness on the basis of eyedness only There are many degrees of native handedness, that is, some individuals are strongly right- or left-handed; efforts to reverse their hand preferences are of no avail and may even be followed by unfortunate consequences Others are so mildly left- or right-handed that they change their preferences with the slightest training Mildly left-handed individuals, because they wish to conform to the majority and because most appliances are designed for right-handed individuals, usually prefer to train their right hands. Children showing consistent right hand preference in a series of tests for handedness and left eye preference in tests for eyedness may well be permitted to establish right hand writing habits. About twenty-three per cent of presumably right-handed children have been found to be left-eyed [1] In regard to left-handedness associated with right-eyedness Ballard[2] found that of fifty-one left-handed individuals fifty-seven per cent proved to be right-eyed, Quinan[3] found fourteen such individuals in a group of twenty-eight. Parsons[4] found only four such cases among six hundred eight right-eyed children These findings made us cautions of hasty diagnosis.

---

[1] Scheidemann, Norma V: A Study of the Handedness of Some Left-handed Writers *The Pedagogical Seminary and Journal of Genetic Psychology*, Vol XXXVII, Dec. 1930, pp 510–16

[2] Gould, G M "Right-handedness and Left-handedness" Lippincott, 1908, pp 18–20

For this study seven of the eight tests used by Haefner[5] in discovering the composition of hand dominance of a group of children, and an additional test, were selected as being suitable for discovering hand preference Each test was given three times, the hand used in two of the trials was recorded as the child's preferred hand [6] In summary the test used for handedness were as follows

1 *Cutting* —The child was required to pick up a pair of scissors placed directly before him and to cut very carefully along an irregular line drawn upon a sheet of paper The hand holding the scissors was recorded

2 *Winding.*—A long cord was fastened to a pencil and placed directly before the child He was then required to wind the cord about the pencil. The hand doing the winding was recorded

3 *Throwing* —The child was required to throw a soft ball to the examiner The hand used to throw was recorded

4 *Receiving.*—The child was given a small object The hand with which the object was received, was recorded

5. *Easy Reaching* —The child was required to reach for a ball placed directly before him on a table The hand used was recorded

---

[1] Mills, Lloyd Eyedness and Handedness *American Journal of Ophthalmology,* Vols I–IX, 1925, pp 106–113 In this study one hundred eighty left-eyed children were found among seven hundred eighty-four presumably right-handed children.

Parsons, B S "Lefthandedness." Macmillan, 1924 Parsons claims that about thirty per cent of all individuals are left-eyed and would be left-handed were it not for the operation of social pressure

[2] Ballard, P B Sinistrality and Speech *Journal of Experimental Pedagogy,* Vol I, 1911–1912, p 298

[3] Quinan, C : A Study of Sinistrality and Muscle Coordination in Musicians, Iron-workers and Others *Archives of Neurology and Psychiatrics,* Vol. VII, 1922, pp 352–300

[4] Parsons *Op cit*, p. 107

[5] Haefner, Ralph "The Educational Significance of Left-handedness" Teachers College Contributions to Education, No 360, 1929

[6] This method of scoring was used by Rife, J M Types of Dextrality. *Psychological Review,* Vol. XXIX, 1922, pp. 474–480

6 *Energetic Reaching*.—The child was required to reach for a ball placed at a distance requiring an energetic reach. The hand used was recorded

7. *Thumb Up*.—Each child was asked to fold his hands by interclasping his fingers The thumb that was placed on top was recorded

8 *Batting*.—The child was required to hold a bat ready to strike a ball about to be pitched by the examiner The hand that was nearer the batting end of the bat was recorded.

The test for eyedness was as follows:

9 *Eyedness.*—A sheet of paper with a small hole torn in the center was placed in the child's hands at about a foot or a foot and a half from his face. A bit of crumpled paper was placed upon the floor The child was directed to look through the hole at the bit of crumpled paper Without permitting the child to move the sheet of paper, the examiner placed her hands alternately over the child's left and right eyes The child reported whether or not he was able to see the bit of paper when one of his eyes was covered Failure to see the bit of paper, indicated that the dominant eye was covered

These nine tests were given to the sixteen left-handed writers. The following table gives the results of the test findings.

HAND- AND EYE-PREFERENCES OF SIXTEEN LEFT-HANDED WRITERS OF A SECOND GRADE GROUP OF THIRTY-FOUR NORMAL CHILDREN

Test	Ervin	Nancy	Carl	Shirley	Marvin	Yvonne	Vera May	Richard	Roberta	Jack	Melville	Raymond	Cecile	Mary	Anita	Jean
Cutting . . .	R	L	R	L	R	R	R	R	L	L	R	L	R	L	R	R
Winding	R	R	R	L	R	L	R	L	L	L	L	L	R	L	R	R
Throwing. . . .	R	L	L	R	R	R	R	R	L	L	L	L	R	R	R	R
Receiving . .	R	L	R	R	R	R	R	R	L	R	L	R	R	R	R	R
Easy reaching	R	L	R	R	R	R	R	R	L	L	L	L	R	R	R	R
Energetic reaching	R	L	R	R	R	R	R	R	L	L	L	L	R	R	R	R
Thumb up	R	L	R	R	R	R	R	R	L	L	R	L	R	L	L	R
Batting .	R	R	R	R	L	R	R	R	L	L	R	L	R	L	L	L
Eyedness	R	L	L	R	R	R	R	R	L	L	R	L	R	R	R	L

According to the test findings, we felt confident that ten of these sixteen left-handed writers should be using their right hands for writing, namely: Ervin, Shirley, Marvin, Yvonne, Vera May, Richard, Melville, Cecile, Mary, and Anita. Carl and Jean, were, perhaps, very mildly left-handed and under other conditions would undoubtedly have established right-handed writing habits.

Before any procedure to reverse the handedness of any child was begun, the second grade teacher called the mothers of the children to the school individually The situation was explained and the series of nine tests was given to the child in the mother's presence so that the mother could see her child's hand preference in acts other than writing In all cases the mothers were very much interested and in all but one case, were eager to cooperate in reversing the child's handedness in writing.

The particular methods employed in effecting a reversal of handedness were different for each child Some children needed scarcely more than an explanation and some enthusiasm over the early right-handed writing attempts. Other children required more encouragement, frequent reminding, and general procedures for training the right hand much like those given to beginning writers In many cases the first attempts with the right hand were better than the child's left-handed writing Most of these children were using their right hands easily and naturally within two or three weeks, and all but one were doing so at the end of the semester (after a period of six weeks). A few weeks after the opening of the fall semester, the teachers of the children whose handedness had been reversed were visited and questioned in regard to the handedness of the children All the children, but the one who had had difficulty in the spring semester, were using their right hands.

During each conference with the mother and later during a questioning of the child, an effort was made to determine the factors that might have influenced the child in using the left hand for writing. The immediate influences for left hand writing in the ten cases definitely right-handed and the two cases who could use their right hands to advantage, seemed to be as follows:

Influenced by left-handedness of	
First grade teacher	1
Mother	1
Playmates or class-mates	4
Statement of second grade teacher	1
Influence not known	5

It is of interest to note that in four cases the mothers did not know that their children were writing with the left hands. The slight incidents that caused some to use their left hands is noteworthy Thus, when the second grade teacher remarked: "We seem to have an epidemic of left hands," one boy decided he, too, would use his left hand "just to make one more." Another child said she began to use her left hand because her playmate broke her right arm and found it hard to use her left hand for writing. The child said she did not want "to be caught like that."

Although the left-handedness of the first grade teacher was found to be a direct influence in establishing but one case of left-handed writing, still we may safely assume that indirectly the teacher's left-handedness played a very important rôle in establishing left-handed writing habits among these children. A left-handed teacher, especially, if she had unhappy experiences or difficulty in trying to establish right-handed writing habits, is, perhaps, more apt to permit a child to establish left-handed habits than is a right-handed teacher Left-handed teachers may be prejudiced toward having a child use his left hand just as right-handed teachers may be prejudiced toward having a child use his right hand.

THE SIGNIFICANCE OF A DIFFERENCE BETWEEN "MATCHED" GROUPS

E. F LINDQUIST

State University of Iowa

A type of experimentation very frequently employed in education and psychology is that which makes use of what are commonly known as "matched groups," or as "matched control groups" It is the purpose of this article to draw attention to an important error in statistical analysis that has been almost universally characteristic of the reports of such experiments, and to suggest an improved statistical procedure and discuss its possibilities

In order to remove any possible ambiguity that may surround the term, "matched groups," it may be well to begin with a specific illustration. Let us consider an experiment to determine the relative effectiveness of the "additive" and the "take-away" methods of teaching subtraction in third grade arithmetic The usual procedure in such an experiment would consist, briefly, of teaching one group of pupils by the "additive" method and another group by the "take-away" method and then of securing, after the period of instruction, a final measure of ability in subtraction for the pupils in both groups. The method then considered the better would be that under which the pupils showed the higher average final ability in subtraction

In order that this procedure be most valid, it is essential that all factors influencing the final measure, except the one factor under investigation—that of method, be kept as nearly as possible the same for both groups One of these factors most important to control is that of the initial ability of the pupils to profit by instruction. To control this factor, it is customary to "match" the two groups with respect to some measure of initial ability This is done by so selecting the two groups that for each pupil in the first group there exists a corresponding pupil in the second group identical with him with respect to this initial measure. In the case of this specific illustration, this initial measure might have been the score on a general intelligence test, or the score on a survey test in arithmetic, or some other similar measure. If the intelligence test score had been used, the two groups would have been so selected as to have identical distributions of intelligence test scores—in other words, they would have been "matched" on the basis of intelligence.

Now it is a well known fact that, even though the two methods are equally effective in general (*i e.*, for the entire population of Grade III pupils), a difference in the average final scores of the two matched groups would almost invariably be found in a *single* experiment of this kind. Assuming equal "true" effectiveness, this would nevertheless occur because the experiment dealt with limited samples, and because differences due to chance in the selection of the samples would alone account for different performances on the final test in subtraction It is therefore the responsibility of the experimenter, before drawing any general conclusions from a single obtained difference, to demonstrate objectively that the difference he obtained was larger than could reasonably be accounted for by chance, or sampling, errors. To do this he must know how large is the expected value of the difference that chance alone would account for in an experiment of this kind. In other words, he must know the Probable Error of the Difference.

In nearly all reports that have yet been made of experiments of this type, the formula that has been employed to determine the value of this probable error has been the familiar PE_{diff} formula

$$PE_{diff} = \sqrt{PE_1^2 + PE_2^2} \qquad (1)$$

in which the PE's under the radical are those of the respective means of the final scores for the two matched groups, each of which have been found through the use of the formula

$$PE_{AM} = 6745\frac{\sigma}{\sqrt{N}} \qquad (2)$$

This Formula for the Probable Error of the Mean Is Not Valid for Use with Matched Groups.[1]—It is based upon the assumption that the samples used are *strictly random* selections from the populations they represent This assumption is not applicable to matched groups. The process of matching on the basis of a measure which is correlated with the final measure *destroys* the randomness of the samples with respect to this final measure The probable amount of sampling error in the obtained difference, instead of being as large as that indicated by the formulas given above, is usually considerably less, in some cases by more than fifty per cent.

The reasonableness of these last two statements may be made more clear by the following illustration, based in part upon the one previously used. Suppose that a large number of samples, each of the same size, and each strictly random (not matched), were selected from the entire population of Grade III pupils. Suppose the pupils in each sample were taught under identical conditions by the *same* method, and then measured by the final test in subtraction Variations in sampling would then result in a variation in the values of the mean scores of these samples on the final test. The standard deviation of these mean scores would be the standard error of any one such mean and, in this case, since random sampling is specified, this standard error could be validly measured by the usual formula for the standard error of the mean.

---

[1] Except in the very rare case where the measures used for matching show no correlation with the final measures

Now let us consider a similar case, but one in which "matching" is involved. If, in the first case (of random sampling) an intelligence test had been given to the same samples, a variation would also have been found in the mean scores on the intelligence test, again due to sampling errors. Now since intelligence is correlated with scores on the subtraction test, a sample that due to chance showed a relatively high mean score on the intelligence test would also show a relatively high mean score in subtraction. If, therefore, a change were arbitrarily made in that sample so as to bring the mean intelligence score down to the mean of all samples, the effect would be to bring the mean subtraction score of that same sample nearer to the mean subtraction score of all samples than it was previously (in the case of random sampling). A restriction of the variation in mean intelligence scores to zero (the effect of identical matching) therefore has the effect of decreasing the chance fluctuations in the mean subtraction scores In other words, the effect of matching is to reduce the value of the standard error of the mean subtraction score of a sample. The amount of this reduction would depend upon the degree of relationship between intelligence scores and subtraction scores.

What is needed, therefore, is a new formula for the standard error of the mean of matched samples that takes into account this restrictive influence of matching upon chance fluctuations in the mean. The type of reasoning already indicated led the author to suggest the problem of deriving such a formula to Mr Samuel S. Wilks, a graduate student in mathematical statistics at the State University of Iowa As a result the desired formula was obtained, and is as follows:

$$\sigma_{M'} = \frac{\sigma\sqrt{1 - r^2}}{\sqrt{N}} \tag{3}$$

in which $\sigma_{M'}$ represents the standard error of means of matched groups. It will be noted that this formula differs from the usual formula only in that the right-hand member is multiplied by the radical $\sqrt{1 - r^2}$, in which r is the Pearson product-moment coefficient of correlation between the measures used as the basis for matching and the measures for which the mean is computed. In the case of the illustration, r is the correlation between intelligence test scores and subtraction scores

Since the original derivation of the formula by Mr Wilks, the author has devised a simpler form of proof, which is reproduced here. This proof is not mathematically so rigid as that provided by Mr. Wilks, but it will perhaps prove easier to understand for the reader not highly trained in mathematical statistics.[1] For the sake of further simplification, this proof will be here set in terms of the specific situation already used in the previous illustrations.

Suppose that we have, for a very large number of pupils selected at random from the Grade III population, the scores on an initial intelligence test and the scores on a final subtraction test given after a course of instruction by one of the two methods described From this large group it would of course be possible to select a number of pupils all of whom made exactly the same score on the intelligence test These pupils would then be "matched" on the basis of intelligence. It would then be possible to predict, without direct calculation, the standard deviation of the scores of these selected pupils on the subtraction test. This could be done through the use of the familiar standard error of estimate formula

$$\sigma_{s'} = \sigma_s\sqrt{1 - r_{is}} \tag{4}$$

in which $\sigma_{s'}$ represents the standard deviation of subtraction scores for the selected pupils, σ_s represents the standard deviation of subtraction scores for all pupils, and in which r_{is} represents the correlation between intelligence scores and subtraction scores

Now, instead of thinking of the scores of individual pupils, suppose that we have, for a very large number of *samples*, all of the same size and all selected strictly at random from the Grade III population, the *mean* scores on the intelligence test and on the subtraction test. Again, as in the case of individual pupils, it would be possible to select a number of *samples* all showing the same *mean* score in intelligence.

---

[1] The original, and more rigid mathematical derivation of the formula will be found in an article by Mr Wilks in this copy of the JOURNAL.

These samples would then be "matched" with respect to intelligence. Again it would be possible, through the use of the standard error of estimate formula, to predict for these selected samples the standard deviation of mean subtraction scores This time, however, formula (4) becomes

$$\sigma_{M_s'} = \sigma_{M_s}\sqrt{1 - r^2_{M_iM_s}} \tag{5}$$

in which $\sigma_{M_s'}$ represents the standard deviation in the mean subtraction scores of the selected (matched) samples, in which σ_{M_s} represents the standard deviation in mean subtraction scores for *all* samples, and in which $r_{M_iM_s}$ represents the coefficient of correlation between *mean* scores in intelligence and *mean* scores in subtraction for all random samples

Now the standard deviation in mean subtraction scores for all random samples is given by the usual formula for the standard error of a mean

$$\sigma_{M_s} = \frac{\sigma_s}{\sqrt{N}} \tag{6}$$

in which σ_s represents the standard deviation in the subtraction scores of individual pupils in a sample, and in which N represents the number of pupils in the sample.

The coefficient of correlation between the means of a series of samples has been shown by Kelley[1] to be equal to the coefficient of correlation between the scores on which the means are based Hence, in this case,

$$r_{M_iM_s} = r_{is} \tag{7}$$

in which r_{is} represents the correlation between intelligence test scores and subtraction scores

By substituting from (6) and (7) in (5) we secure the formula

$$\sigma_{M_s'} = \frac{\sigma_s\sqrt{1 - r^2_{is}}}{\sqrt{N}} \tag{8}$$

which, by dropping the subscripts peculiar to the illustration, becomes the formula already given as formula (3).

Substituting the value of the standard error of the mean of matched groups as given by formula (3) in the usual formula for the standard error of a difference between means, we secure the following as the

---

[1] Kelley, Truman L, "Statistical Method" P 178, formula (118)

correct formula to employ for determining the standard error of a difference in obtained means for two matched groups

$$\sigma_{\text{diff}} = \sqrt{\left(\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}\right)(1 - r^2)} \tag{9}$$

where σ_1 and σ_2 represent the standard deviations in final scores of the respective groups, where N_1 and N_2 represent the number of cases in each, and where r represents the Pearson product-moment coefficient of correlation between the measures used as the basis for matching and the measures between which the final comparison in means is made.

It is obvious, from an inspection of formula (9), that the true standard error of an obtained difference between matched groups might be considerably less than that yielded by the usual formulas, and that hence many differences which have been considered accountable for by sampling errors might have been statistically significant. Fortunately, the incorrect use of the formulas based on the assumption of random sampling resulted in errors only in the direction of conservativeness. That fact, however, is no valid excuse for their continued application, especially when the correct technique proves so simple in application.

It is important to note that formula (9) does not depend upon the relation of the mean initial scores of the matched groups to the "true" initial score of the entire population which is being considered For example, if the matched samples of the original illustration had due to chance been above or below the average of all Grade III children in intelligence, the formula would still apply This follows by analogy from the fact that the standard error of estimate formula applies to all values of the variable from which the estimates are made.

The proof of formula (9) that is contained in this article also suggests that the formula applies even though the exact distribution of initial scores is not the same for both matched groups, if only their *mean* initial scores are the same. This indicates that the formula should be valid for use with groups that have not been matched "pupil for pupil," but in which the means and standard deviations alone have been equated. A more rigid mathematical proof of this proposition, however, should be provided before much confidence is placed in it.

It is also important to note that formula (9) does not indicate how far the obtained difference between two matched samples is likely to deviate from the difference that would have been obtained had the *entire* population been measured, but tells only how far the obtained difference is likely to deviate from the difference that would have been found between infinitely large groups showing the *same distribution of initial measures as that matched samples that were used* Where the purpose of experimentation, however, is simply to determine (as is usually the case) whether there is or is not *any* real difference between the two methods of learning or teaching (*i e.*, whether or not the obtained difference is statistically significant, and where the absolute amount of the difference for the entire population is not demanded) this latter caution is not of very great practical importance If, for example, a method is truly superior for pupils at one level of intelligence, it is certainly likely to be superior for pupils at another level not far removed from the first If reasonably large samples are used ($N > 30$), and if the first sample is selected roughly at random from the entire population and the second matched *with* it (a usual procedure), it is not reasonable to suppose that the difference in intelligence between the matched groups and the entire population (due to sampling error in the first sample selected) will be large enough to have any effect upon the validity of conclusions concerning the general relative effectiveness of the methods investigated

It should hardly be necessary to point out that formula (3) is useful only to measure sampling fluctuations from sample to sample in the means of samples that have been matched with respect to a related variable It in no sense, therefore, supplants the usual formula for the standard error of a mean, since that formula is in no case valid in the specific situation to which formula (3) applies. As far as most practical applications are concerned, formula (3) need not concern the investigator at all, formula (9) being the only one that need be directly applied in the usual "matching" experiment In the case, however, where the effect of the methods compared is to result in different values of the correlation coefficient between initial and final measures for the two groups, it is necessary to apply formula (3) independently to the mean of each sample, and then substitute the resulting values for the standard errors of the means in the usual formula for the standard error of the difference Formula (9) is given in combination form for the situation where any difference in method does not seriously affect the correlation between initial and final scores, which condition usually exists Even though the value of r is slightly different for the two groups, formula (9) will still yield a useful and close approximation to the true standard error of the difference. It is considered desirable to propose formula (9) in combination form since it eliminates the necessity of calculating the standard error of each mean independently, and hence avoids the danger that the less critical investigator will attempt to interpret these standard errors in the way in which the usual standard error of the mean is interpreted.

THE STANDARD ERROR OF THE MEANS OF "MATCHED" SAMPLES

SAMUEL S WILKS

State University of Iowa

The problem of determining the variation in the mean of one character of the items of a sample when the distribution of another correlated character is made identical for all samples, item by item, with an arbitrary distribution, was suggested to me by Prof. E F Lindquist of the State University of Iowa A discussion of the use and importance of the theory as a statistical technique in certain types of experimental work will be found in an article by Lindquist in this issue of the JOURNAL In this paper I shall consider only the mathematical derivation of the expression for this variation.

In order to state the problem more accurately, we may describe the type of sampling involved in the following manner. Suppose a sample of N items is drawn from a population in such a way that the distribution of a particular character X of each item is made identical, item for item, with a given X-distribution. This selection or matching of X's is made at random relative to a second character Y and the only case of general interest is that in which there is a correlation between X and Y The question naturally arises: What is the expected variation in the mean[1] of the Y's due to such sampling? It should be noted that the given distribution of X may be of an arbitrarily selected form or it may be derived at random The results, as we shall see, are independent of the form of this distribution

Let us assume that X and Y are normally distributed, and that a correlation of r exists between them As was previously stated, we can obtain more general results at once without making this assumption of normality, but it is the first one made because of the greater practical interest in the case of a normal distribution For convenience, let us think of the X-Y plane as being divided into small squares Then in any finite sample of N items there are only a finite number of squares into which the points (x, y) (each representing an item of the sample) will fall. The distribution of Y's associated with a given value x_p of x which is taken as the mid-point of the p-th interval of X's is called the x_p array of Y's. Similarly y_i is the midpoint of the i-th interval of Y's. Of the N pairs of X and Y, n_p will have the value of the X character in the interval x_p which form the x_p array of Y's. This array will have its mean and its standard deviation, which we will denote by $\bar{y}_p$ and σ_p respectively. The mean of all Y characters will be $\bar{y}$ and their variability will be given by $\sigma_{\bar{y}}$. The number of the N items falling into the square with center (x_p, y_i) will be denoted by n_{pi}.

---

[1] I have obtained expressions for the variation in the standard deviation of Y and in the correlation coefficient r between X and Y, but have not included these in this article. They will be made available later in my Doctor's thesis

Let δ denote the sampling deviation of a variate from its true value. Then since the distribution of X's is constant from sample to sample, it follows that the deviation δn_p is zero. We shall make use of the propositions that if s and s' are the frequencies in any two mutually exclusive non-independent (when either or both are subjected to random sampling variation) frequency classes of a frequency distribution of M elements the standard deviation of s due to random sampling is given by

$$E(\delta_s{}^2) = \sigma_s{}^2 = s\left(1 - \frac{s}{M}\right) \tag{1}$$

and the correlation $r_{ss'}$ between deviations in s and s' due to random sampling is given by

$$E(\delta s \quad \delta s') = r_{ss'}\sigma_s\sigma_{s'} = -\frac{ss'}{M}. \qquad (2)^1$$

Now it is obvious that we may write

$$N\bar{y} = S_p[S_i(n_{pi}y_i)] \tag{3}$$

where the sums are taken over all values of p and i.

Taking the variation we have

$$N\delta\bar{y} = S_p[S_i(\delta n_{pi}y_i)] \tag{4}$$

squaring

$$N^2(\delta\bar{y})^2 = \left\{S_p[S_i(\delta n_{pi}y_i)]\right\}^2$$
$$= S_p[S_i(\delta n_{pi}y_i)]^2 + S_{pp'}[S_i(\delta n_{pi}y_i)S_j(\delta n_{p'j}y_j)] \tag{5}$$

where $S_{pp'}$ denotes the summation over all values of p and p' except for $p = p'$.

Expanding again,

$$N^2(\delta\bar{y})^2 = S_p[S_i(\delta n^2{}_{pi}y_i{}^2)] + S_p[S_{ij}(\delta n_{pi}\delta n_{pj}y_iy_j)]$$
$$+ S_{pp'}[S_i(\delta n_{pi}\delta n_{p'i}y_i{}^2)] + S_{pp'}[S_{ij}(\delta n_{pi}\delta n_{p'j}y_iy_j)] \tag{6}$$

where the meaning of S_{ij} is obvious.

---

[1] For proof, see Rietz, H. L.: "Mathematical Statistics," 1927, p. 119.

Summing and dividing by the number of possible samples of this kind, and using the proposition that the expected value of a sum is equal to the sum of the expected values, we get

$$N^2\sigma_{\bar{y}}^2 = S_p[S_i(E(\delta\bar{n}_{pi}^2)y_i^2)] + S_p[S_{ij}(E(\delta n_{pi}\delta n_{pj})y_iy_j)]$$
$$+ S_{pp'}[S_i(E(\delta n_{pi}\delta n_{p'i})y_i^2)] + S_{pp'}[S_{ij}(E(\delta n_{pi}\delta n_{p'j})y_iy_j)]. \quad (7)$$

But from (1) it is evident that

$$S_p[S_i(E(\delta\bar{n}_{pi}^2)y_i^2)] = S_p\left[S_i\left(n_{pi}\left(1 - \frac{n_{pi}}{n_p}\right)y_i^2\right)\right] \quad (8)$$

and by (2)

$$S_p[S_{ij}(E(\delta n_{pi}\delta n_{pj})y_iy_j)] = -S_p\left[S_{ij}\left(\frac{n_{pi}n_{pj}}{n_p}\right)y_iy_j\right]. \quad (9)$$

Let us consider the expression $E(\delta n_{pi}\delta n_{p'j})$ for p different from p' Since n_p is invariable from sample to sample, it follows that the deviation δn_{pi} for any i within the group n_p cannot affect the deviation δn_p which is zero and therefore cannot affect the deviation $\delta n_{p'i}$ of the sub-group $n_{p'i}$ in any other group $n_{p'}$ for p different from p'. Hence the deviations δn_{pi} and $\delta n_{p'j}$ are independent for all values of p and p' provided p is different from p'

Hence

$$E(\delta n_{pi}\delta n_{p'j}) = [E(\delta n_{pi}) \cdot E(\delta n_{p'j})]. \quad (10)$$

Thus the last two sets of terms of (7) vanish, and combining (8) and (9) and placing this value in (7), we get

$$\begin{aligned} N^2\sigma_{\bar{y}}^2 &= S_p\left\{[S_i(n_{pi}y_i^2)] - \frac{[S_i(n_{pi}y_i)]^2}{n_p}\right\} \\ &= S_p[S_i(n_{pi}y_i^2) - n_p\bar{y}_p^2] \\ &= S_p[S_i(n_{pi}(y_i - \bar{y}_p)^2)]. \end{aligned} \quad (11)$$

But

$$S_i[n_{pi}(y_i - \bar{y}_p)^2] = n_p\sigma_p^2$$

hence

$$N^2\sigma_{\bar{y}}^2 = S_p[n_p\sigma_p^2] \quad (12)$$

Since we have assumed normal distribution, and thus a homoscedastic system with linear regression, all arrays of Y's have the same standard deviation $\sigma_y\sqrt{1 - r^2}$.

Then

$$S_p[n_p\sigma_p^2] = Ns_y^2 = N\sigma_y^2(1 - r^2)$$

The standard error of the mean $\bar{y}$ is finally

$$\sigma_{\bar{y}} = \frac{\sigma_y\sqrt{1 - r^2}}{\sqrt{N}} \tag{13}$$

which we note again is independent of the distribution of X in the sample. It must not be construed, however, that the expected value of the mean of the Y's is independent of the distribution of the X's. I state, without giving proof here, that the expected value of $\bar{y}$ is given by

$$\bar{y} = y_0 + r\frac{\sigma_y}{\sigma_x}(\bar{x} - x_0) \tag{14}$$

where x_0 and y_0 are the true means of the X's and Y's for the entire population and $\bar{x}$ is the mean of the given distribution of X's

By similar methods it is not difficult to show that if all of the items of a sample are selected or matched on $s - 1$ characters, where s characters are normally distributed and not independent, and with linear multiple regression, then the standard error of the means of the character X_s is given by

$$\sigma_{\bar{s}} = \frac{\sigma_s\sqrt{1 - r_{s.12\ldots s-1}}}{\sqrt{N}}\,. \tag{15}$$

where $r_{s.12\ldots s-1}$ is the multiple correlation coefficient of order $s - 1$ of X_s with the $s - 1$ variables

It may be of interest to know that all of the results contained in this paper have been confirmed independently through the application of an entirely different and somewhat more rigorous method of proof, which will be made available in a thesis soon to be published at the State University of Iowa

RELIABILITY OF INTEGRATION INDEX DIFFERENCES

JOHN W DICKEY

State Normal School, Newark, N J

The formula

$$K = \frac{M}{\sigma} \tag{1}$$

has been presented,[1] by the writer, as a quantitative measure of pupil integration within the public schools The formula giving the reliability for separate indices has also been reported [2]

The necessary comparison of indices (whether made by the same population on different tests or on different forms of the same test, or made by uncorrelated populations on different tests or the same test) is statistically impossible without the use of formulas which yield the reliability of their differences.

It is the purpose of this paper to derive the standard error and the probable error formulas which are necessary in the comparison of indices for both the correlated and the uncorrelated groups

Let,

M_1 = mean gross score on test one
M_2 = mean gross score on test two
σ_1 = standard deviation of gross scores on test one
σ_2 = standard deviation of gross scores on test two
r_{12} = correlation between test one and test two
N = the population
$K_1 = \frac{M_1}{\sigma_1}$ = Integration Index on test one
$K_2 = \frac{M_2}{\sigma_2}$ = Integration Index on test two
$\Delta = K_2 - K_1$ = an index difference

The standard deviation of such index differences, as

$$\Delta = \frac{M_2}{\sigma_2} - \frac{M_1}{\sigma_1} \tag{2}$$

is the required formula

---

[1] An Index of Integration. *Journal of Educational Psychology*, Vol XX, No 9, Dec, 1929, p 625

[2] Note on the Reliability of the Index of Integration *Journal of Educational Psychology*, Vol XXI, No 3, March, 1930, p 231

By writing the total derivative of equation (2), we obtain the differential equation

$$d\Delta = \frac{\sigma_2 dM_2 - M_2 d\sigma_2}{\sigma_2^2} + \frac{M_1 d\sigma_1 - \sigma_1 dM_1}{\sigma_1^2} \tag{3}$$

When squaring, summing and dividing by the theoretical infinite population, we get the variance of Δ in the form

$$\sigma_\Delta^2 = \frac{\sigma^2{}_2\sigma^2{}_{M_2} + M^2{}_2\sigma^2{}_{\sigma_2} - 0}{\sigma_2^4} + \frac{\sigma^2{}_1\sigma_{M_1} + M^2{}_1\sigma^2{}_{\sigma_1} - 0}{\sigma_1^4}$$
$$+ 2\left(\frac{-M_1M_2\sigma_{\sigma_1}\sigma_{\sigma_2}r_{\sigma_1\sigma_2} - \sigma_1\sigma_2\sigma_{M_1}\sigma_{M_2}r_{M_1M_2} + 0 + 0}{\sigma_1^2\sigma_2^2}\right) \tag{4}$$

The zeros in equation (4) occur because the correlation between the two independent variables, M and σ, is zero.

The following simplified formulas are the result when we substitute[1]

$$r_{M_1M_2} = r_{12},\ r_{\sigma_1\sigma_2} = r^2{}_{12},\ \sigma^2{}_M = \frac{\sigma^2}{N},\ \sigma^2_\sigma = \frac{\sigma^2}{2N}, \text{ and } K = \frac{M}{\sigma},$$

in equation (4) and reduce to a convenient form.

$$\sigma_\Delta^2 = \frac{1}{2N}[4 + K_1^2 + K_2^2 - r_{12}(K_1K_2r_{12} + 2)]$$

whence

$$\sigma_{(K_2-K_1)} = \frac{1}{\sqrt{2N}}\sqrt{4 + K_1^2 + K_2^2 - r_{12}(K_1K_2r_{12} + 2)} \tag{5}$$

and

$$PE_{(K_2-K_1)} = \frac{0.6745}{\sqrt{2N}}\sqrt{4 + K_1^2 + K_2^2 - r_{12}(K_1K_2r_{12} + 2)} \tag{6}$$

In case the reliability of index differences (which occur when different forms of the same test are used) is to be substantiated, the correlation coefficient is at once the reliability coefficient of the test.

If the reliability of index differences (for uncorrelated groups) is to be investigated, formulas (5) and (6) reduce immediately to the formulas

$$\sigma_{(K_2-K_1)} = \frac{1}{\sqrt{2}}\sqrt{\frac{2 + K_1^2}{N_1} + \frac{2 + K_2^2}{N_2}} \tag{7}$$

---

[1] Kelley, T. L.: "Statistical Method." New York The Macmillan Company, 1923, p. 178, formulas 118 and 121. The assumptions, in the case of $r_{\sigma\sigma} = r^2$, of rectilinearity, homoscedasticity, and equal kurtosis, do not vitiate our findings since a variation of forty points in the correlation coefficient is required to create around a ten per cent error in formulas (5) and (6).

and

$$PE_{(K_2-K_1)} = 0.4769\sqrt{\frac{2+K_1^2}{N_1}+\frac{2+K_2^2}{N_2}} \tag{8}$$

where N_1 and N_2 are the respective populations.

Formulas (5) and (6) are used when indices are compared for any given population; whereas, formulas (7) and (8) are used when indices are compared for *entirely* different populations.

PARENTAL AGE AND INTELLIGENCE OF OFFSPRING

MINNIE LOUISE STECKEL

University of Chicago

The object of this study was to disclose, if possible, any relationship which might exist between the intelligence of children and parental ages at the time of the birth of the child on the basis of findings on a large population by the use of intelligence tests.

The data upon which this study is based were obtained from public school children at Sioux City, Iowa and their parents during the school years 1926–1928. The study includes records of children from Grade I to Grade XII inclusive. Intelligence ratings of the children were obtained by means of group intelligence tests. Questionnaires were sent to the parents asking for their own ages, and the ages and birth order of all their children.

Four different group tests were used in order adequately to cover the age range of intelligence. The Kuhlmann-Anderson Test was used for grades I to Junior III; the National Intelligence Test for grades Senior III to Senior IV, the Otis Intermediate Test for grades Junior V to Senior VIII, and the Otis Advanced Test for grades Junior IX to Senior XII. Several thousand children were tested with each test. It was possible, therefore, to restandardize on the same basis the intelligence quotients as obtained by each of the four tests, so that the results of all four tests are directly comparable. The standard score is calculated in the following manner:

$S \frac{(IQ) - M}{\sigma}$ in which S is the standard score, M is the mean intelligence quotient for the year group of the child's age and σ is the standard deviation of the intelligence quotients for that age group. By using the intelligence quotient upon which to base the calculations rather than the raw score, the measure is uncorrelated with age. By transmuting the intelligence quotients of each test into standard scores the possibility of differences which might arise due to the relative difficulty of the tests at the various age levels is eliminated. The constant 5.00 has been added to the standard scores as calculated in order to avoid negative values. Thus a child whose rating is 5.00 has an intelligence rating equal to the average of all children of Sioux City of his age. A child with a score of 3.80 has a standard score of −1.20 for children of his own age group. Only records of normal children of the Caucasian race are included in this study.

Some parents may have stated their own ages incorrectly There was, however, nothing compulsory about answering the questionnaire, therefore these cases would be so few comparatively that they would scarcely affect the validity of the study. In Table I columns C and D show the distribution and mean intelligence of the children for each two year age period of the parents Figure 1 presents graphically the relationship between the mean intelligence of children born during each two year age period of the mothers Curve C shows this relationship for the first-born children, and Curve D for all children regardless of birth-order. Both curves indicate that children born of mothers who are less than approximately twenty-six to twenty-eight years of age are, in general, less intelligent than children born of mothers who are this age or older.

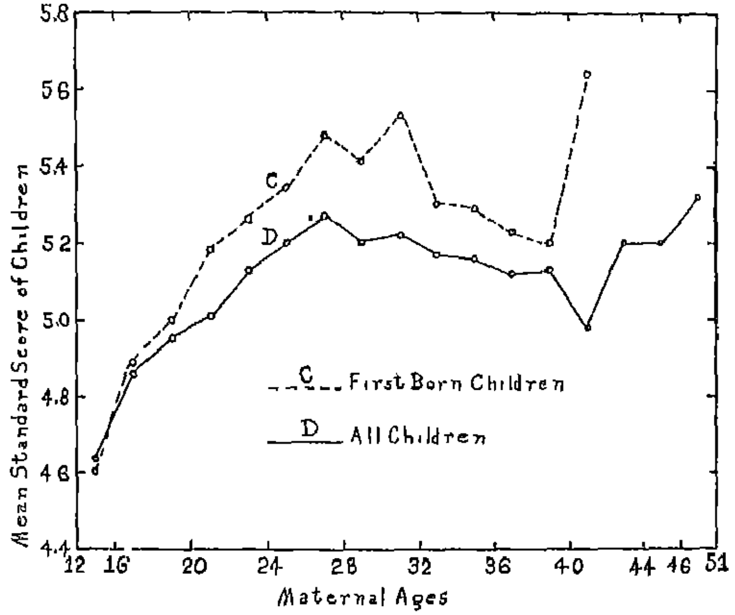


Fig 1.

Curves A and B of Fig. 2 shows a similar relationship between intelligence of children and paternal ages. The chief difference is in the fact that the curves representing intelligence of children and paternal ages indicate that children born of fathers under approximately thirty or thirty-two years of age are in general less intelligent than children born at this paternal age or older, whereas (as has been said) this period with respect to maternal ages extends approximately only to the twenty-sixth to twenty-eighth year.

Although the curves indicate periods of very high intelligence after twenty-eight and thirty-two years for the maternal and paternal ages respectively, these cases are at the extreme of the curve and are represented by a relatively small number of cases. In general, after these age periods, there is a slight drop in the curves representing children's intelligence.

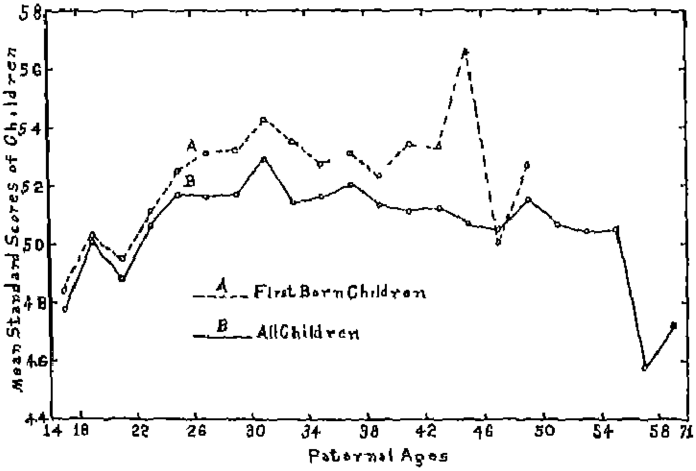


Fig 2

The curves C and A representing intelligence of first-born children and parental ages are consistently higher than curves D and B representing the intelligence of all children regardless of birth-order.

These facts as presented might have several possible interpretations. It may be conjectured that whatever causes the vitality of the human body to increase as maturity is approached and to diminish with advancing age may also affect the uniting reproductive cells and have a deleterious effect upon the mentality of the offspring. Another possible interpretation may be that the extraordinary somatic difficulties accompanying child birth in immature and in elderly mothers might produce so great initial handicaps upon their children that they may never be fully compensated by the time the children reach maturity

It might be that the differences in environmental factors for children of very young or elderly parents as compared to those parents of the intervening age period are pronounced enough to affect to an appreciable degree the rating of children on an intelligence test given at school.

Possibly the explanation of the results of this study is to be found largely on a nationality and socio-economic basis. Grouping parents according to age necessarily involves nationality, and socio-economic groupings also. Although the data are limited to the Caucasian race, either one or both of approximately one-fourth of the parents of the children are foreign-born. These are largely of Russian, Lithuanian, and German extraction. There is a tendency for European stock to marry young. It also is probable that children of immigrants, on the whole, would rate a little lower on an intelligence test than would children of native-born parents.

The army examinations showed that the various occupational groups draw men of different intelligence levels. These occupational groups, therefore, constitute social classes based on intelligence Studies by Hirsh,[1] Kornhauser,[2] and Haggerty and Nash,[3] show that the occupation of the father also is a fair index of the intelligence of his children, *i.e.*, that there is a very real association between parental occupation and intelligence of offspring.

Children of very young parents are children of parents whose occupational choice does not demand college or professional training. Many of these parents do not even go to high school. They drop out of school as soon as attendance laws permit or as soon as labor laws permit them to work. Not that going to college on the part of the parents appreciably increases the intelligence of their offspring but the parents in occupations which do not demand a long period of training are of a stratum of society which lives on a lower economic basis. The more intelligent children are offspring of parents who have gone to college and professional schools. These parents of necessity marry later. Their children are not more intelligent because the parents are older and possibly not because their parents have had more education but rather because their parents come from a higher, more intellectual stratum of society The scale of living of this stratum of society demands that marriage and reproduction be delayed until the parents are economically able to maintain their family on a scale of living to which they themselves have been accustomed.

---

[1] A Study of Natio-racial Mental Differences. *Genetic Psychology Monograph*, 1926, pp. 239–407

[2] The Economic Standing of Parents and the Intelligence of Their Children *Journal Educational Psychology*, Vol LX, 1918

[3] Mental Capacity of Children and Parental Occupation *Journal Educational Psychology*, 1924, pp 559–572

The explanation of the fact that the intelligence of the child correlates as closely with the father's age as with the mother's age (except that the period of greatest intelligence for the child comes three or four years later for the father's age) might also be explained on a socio-economic basis rather than on a hereditary or biological basis accounted for in the fact of the later maturing of the male parent. The demand made on the father in support of his family delays his marriage and his reproductive period several years later than the mother's in order that he may first establish himself economically and socially. That there is the same relationship between intelligence of children and paternal ages as between intelligence of children and maternal ages might be explained in the fact that the paternal ages correlate positively with the maternal ages and therefore bear a similar relationship to the intelligence of their offspring.

The seemingly greater intelligence of first-born children, as indicated by the higher curves A and C, undoubtedly has its explanation on a socio-economic basis. It is generally understood that the lower economic classes have larger families than do the professional classes. The children coming from the lower occupational groups also have lower intelligence ratings As long as we consider only the intelligence of the first-born children, each family is represented only once. The superior intelligence of the children of the professional classes raises the intelligence level of the entire group. In curves B and D, representing the intelligence of all children against parental age, the relatively lower intelligence of the greater number of children from families of lower socio-economic groups lowers the mean intelligence of the entire group. The greater intelligence of children represented in curves A and C is to be explained, then, not in the fact that they are first-born children but in the fact that each family of the lower occupational classes has no more representatives in the group than each family of the professional classes.

If late marriage and reproduction accompany a rising economic status, then the children of elderly parents should exceed the children of middle-aged parents in intelligence A slight drop of the intelligence curve of children of elderly parents might conceivably be due to biological factors such as poor productive stock with a low fecundity-fertility ratio or to the physical decline of the older parent. The seemingly lower intelligence of children of this group, as indicated by the results of this study, might also find its interpretation on a socio-economic basis; for not only do parents of the lower economic classes reproduce at an earlier age than do the professional classes but there is a tendency for child-bearing to be continued throughout the entire reproductive period of the mother of the laboring classes while for the professional classes child bearing is restricted toward the close of the reproductive period as well as at its beginning. A slight decrease in intelligence of children of elderly parents, as indicated in the study, might be accounted for in the fact that this part of the curve again represents a preponderance of children from the lower occupational groups. Undoubtedly this drop in the curve of intelligence would be more pronounced were it not for the fact that within the same family intelligence increases with ordinal number as shown in an earlier study [1]

The relationship between parental ages and intelligence of their children as indicated by the results of this study is doubtless a true relationship. The conclusion that there is a direct causal relationship between parental ages and the intelligence of their children is not at all justified by the results of the present study. The writer makes no claim to having established the cause or causes underlying the results of the study

If the population in question were first classified into occupational groups and the age of parents and intelligence of children compared, any differentiation might indicate biological factors operating which are dependent upon parental age. Present indications are that in such occupational groups the variation in intelligence of the children would be no greater than the occupational variation of the parents within each group

A subject closely associated with parental age and intelligence of offspring is amount of disparity between the ages of the two parents as compared to the intelligence of their offspring Custom and general

---

[1] Steckel, Minnie Louise: Intelligence and Birth Order in Family *Journal of Social Psychology*, August, 1930

TABLE I.—PARENTAL AGE AND INTELLIGENCE OF CHILDREN

A. First-born children			B. All children			C. First-born children			D. All children			Disparity between parental ages	E. First-born children		F. All children	
Fathers' ages	Mean scores of children	Total number of children	Fathers' ages	Mean scores of children	Total number of children	Mothers' ages	Mean scores of children	Total number of children	Mothers' ages	Mean scores of children	Total number of children		Mean scores of children	Total number of children	Mean scores of children	Total number of children
14–17	4 85	17	14–17	4 77	21	12–15	4 60	29	12–15	4 64	34	−13 to −7	5 51	13	5 20	35
18–19	5 03	101	18–19	5 01	113	16–17	4 89	179	16–17	4 86	208	− 6 and −5	5 27	23	4 90	68
20–21	4 95	333	20–21	4 88	410	18–19	5 00	539	18–19	4 95	674	− 4 and −3	5 21	66	5.12	185
22–23	5.12	508	22–23	5 06	779	20–21	5 18	704	20–21	5 01	1,146	− 2 and −1	5 28	198	5 10	584
24–25	5 25	620	24–25	5 17	1,091	22–23	5 26	643	22–23	5 13	1,393	0	5 32	463	5 21	1,154
26–27	5 31	577	26–27	5 16	1,258	24–25	5 34	572	24–25	5 20	1,432	1 and 2	5 28	842	5 22	2,321
28–29	5 32	455	28–29	5 17	1,231	26–27	5 48	350	26–27	5 27	1,239	3 and 4	5 25	707	5 15	1,983
30–31	5 43	342	30–31	5 29	1,149	28–29	5.41	267	28–29	5 20	1 114	5 and 6	5 23	540	5 16	1,670
32–33	5 35	214	32–33	5 14	975	30–31	5 53	158	30–31	5 22	864	7 and 8	5 20	318	5 14	963
34–35	5 27	172	34–35	5 16	862	32–33	5 30	97	32–33	5 17	753	9 and 10	5 20	236	5 11	760
36–37	5 31	107	36–37	5 20	706	34–35	5.29	49	34–35	5 16	587	11 and 12	5 17	103	5 12	365
38–39	5 23	86	38–39	5 13	566	36–37	5 23	35	36–37	5 12	451	13 and 14	5 16	63	5 06	173
40–41	5 34	46	40–41	5 11	431	38–39	5 20	22	38–39	5 13	326	15 and 16	4 85	34	4 88	136
42–43	5 33	27	42–43	5 12	335	40–51	5.64	23	40–41	4 98	199	17 and 18	5 33	23	4 90	76
44–45	5 66	24	44–45	5 07	227	.	.		42–43	5 20	101	19 and 20	5 28	19	5 09	64
46–47	5 00	10	46–47	5 05	155		..	..	44–45	5 20	44	21 to 34	4 78	12	4 70	37
48–57	5 27	21	48–49	5 15	108		.		46–51	5 32	17					
			50–51	5 07	54											
			52–53	5 04	46											
			54–55	5 05	23											
			56–57	4 58	17											
			58–71	4 72	23											
	Total, 3,660			Total, 10,580			Total, 3,665			Total, 10,582			Total, 3,660		Total, 10 574	

opinion presumably would favor the father being from two to five years older than the mother. In Table I, columns E and F give the distribution and mean intelligence of the children for each two years disparity between the ages of the two parents. This relationship is shown graphically in Fig. 3 Curve E represents the mean intelligence of first-born children and curve F represents the intelligence of all children against the disparity between parental ages The disparity between the ages of the two parents is indicated by the "Fathers' age Minus the Mother's age." When the father is the older parent the difference between the ages is indicated as +, when the mother is older the difference is indicated as −; and 0, of course, indicates that both parents are of the same age. Considering both curves E and F, the mean intelligence of the children decreases as the disparity between the parental ages increases. The curves both show a more rapid decrease in the intelligence of the children when the mother is the older parent than when the father is older At the extremes, in both + and − directions, the cases are so few that the curve is very irregular However, the general tendency downward, indicating lower intelligence, is quite apparent. The curve of intelligence of the first-born children against disparity of parental age is higher than when the intelligence of all children is represented. Here again,

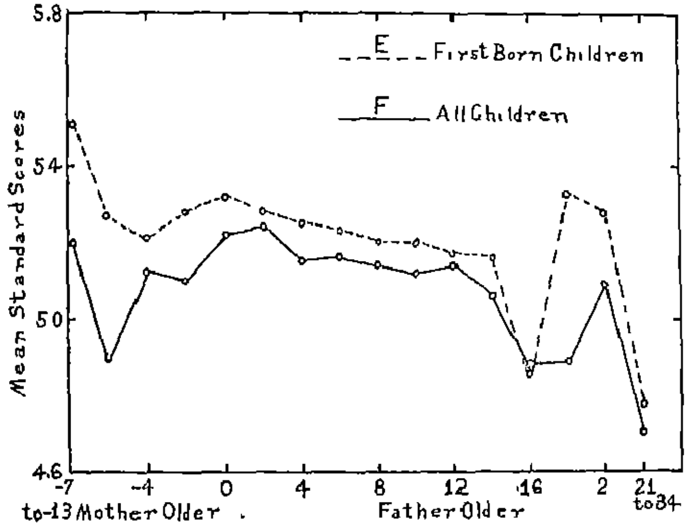


Fig 3

as with parental age and intelligence, the larger families having more representatives and lower intelligence ratings, lower the mean intelligence of the entire group. When the mean intelligence of only the first-born children is considered, the families of the professional classes have equal representation with the lower occupational groups, so the mean intelligence of the group is higher

The writer knows of no study indicating whether greater disparity between parental ages is more common between foreign born parents or between native-born-parents, between parents of the professional classes or between parents of the lower socio-economic groups, or in which groups the mother more often is the older. Such evidence and an occupational classification is necessary before the causes which are operative to produce the relationship between intelligence of child and amount of disparity between parental ages can be determined.

The results of this part of the study indicate that the greater the disparity between parental ages the less favorable is the prognosis for the intelligence of the offspring The prognosis, however, is more favorable if the father is the older parent than if the mother is the older parent if the amount of disparity between the ages of the two parents is greater than four or five years

CONCLUSIONS

In general, as shown by intelligence tests, children born of very young parents are less intelligent than children born of more mature parents Below the age of twenty-six to twenty-eight for mothers and thirty to thirty-two years for fathers, the younger the parents the less favorable is the prognosis for the intelligence of the offspring.

The present study indicates that the nearer the ages of the two parents approach each other the more favorable is the prognosis for the intelligence of their children. The prognosis for the intelligence of the children is less favorable as the disparity between parental ages grows extreme When extreme disparity exists between parental ages the prognosis for the intelligence of the child is better if the father is the older parent than if the mother is the older parent

The writer presents several possible interpretations of the results of the study but makes no claim to having established the factors which might be operative in producing the results as presented. Further study must be made to reveal which interpretation most nearly approaches the truth.

SEX DIFFERENCES: COLLECTING INTERESTS

PAUL A. WITTY

Northwestern University

AND

HARVEY C. LEHMAN

Ohio University

In a previous article the writers have questioned the instinctiveness of the collecting tendency.[1] They have mentioned also some of the difficulties that one encounters when he attempts to define the term "collecting." Definition of "collecting" appears to be varied and inconsistent; it is nevertheless true that studies of the articles which children report that they actively collect may reveal certain intrinsic interests of children. Therefore, it seems logical that such interests might be used profitably in motivating school work. Collecting interests appear to be associated intimately with the growing self; the recognized identity of the interest and the growing self (with consequent inner urge) should provide a genuine actuator of interest and consequent success in school work. Of course, numerous interests are undesirable; these should be recognized, sublimated, and redirected. Others, however, seem salutary manifestations of growth These should have abundant opportunity for expression in the curricular activities Genetic studies of behavior manifestations seem to the writers to be of inestimable value in guiding and developing school children

In a previous article,[1] the writers presented a list of approximately 200 articles which children listed as ones they were actively collecting in 1927–1928 Further analysis of these data has yielded certain indications of sex differences in collecting interests which appear to be significant

As a means of studying the sex differences in collecting interests, the writers listed the articles* that were collected more than twice as commonly by boys as by girls, they listed also the articles that were collected more than twice as frequently by girls as by boys By this means two lists were obtained. The first list therefore includes only those collections which seem to appeal predominantly to girls; the second list includes only those collections which appeal strongly to boys.

---

* These data were assembled by one of the writers and several graduate students The table containing the articles most frequently collected, and the method of investigation may be obtained by reference to this JOURNAL, Vol XXI, 1930, pp 112–128

The articles which are collected much more frequently by girls than by boys appear in Table I, those which have a much more intense appeal to boys than to girls are listed in Table II.

In Table I, the articles collected by the girls are classified under seven headings, namely: (*a*) Objects possessing æsthetic appeal or value, (*b*) objects for personal adornment, (*c*) objects of sentimental appeal, (*d*) dolls, doll clothes, etc , (*e*) household accoutrements, (*f*) souvenirs (predominantly from the classroom), and (*g*) objects used in playing games. In Table II, the articles of unusual interest to the boys are assembled under six headings, namely: (*a*) Animal parts and insects, (*b*) junk (to sell), (*c*) tobacco souvenirs, (*d*) objects associated with war, hunting, fishing, etc , (*e*) objects used in playing games, and (*f*) miscellaneous ones.

In examining Tables I and II certain facts should be kept in mind. In the first place one should recall that this presentation seeks to place in sharp relief *sex differences*, not sex likenesses. If this fact be overlooked, the reader might exaggerate the significance of the sex differences. For example, in Table I (*a*), the girls are shown to have collected more often than the boys objects of æsthetic appeal or value. Nevertheless, one or more boys collected all but one of the items listed in part (*a*) of Table I. The items listed in Table I (*a*) are not articles collected only by girls; they are items which were collected more than twice as frequently by girls as by boys.*

One must bear in mind also that these data were obtained within a rather restricted geographical area. A specific environmental background is therefore reflected. For example, in Table II (*a*) it will be found that the boys collected furs, rabbits' ears, gopher skins, etc It is clear that city boys' collections would yield no such collections. For this reason the sex differences herein reported should be interpreted with the realization that they obtain for a specific locality only Nevertheless, when these data are viewed from the standpoint of the several *groupings* rather than from the standpoint of specific items, it seems probable that the sex differences may be representative of widespread and rather general tendencies.

---

* The writers have previously discussed the fact of sex differences in æsthetic appreciation. (See *American Journal of Psychology*, Vol XL, July, 1928, pp 449–457.) The present findings corroborate the conclusions set forth previously.

TABLE I.—ITEMS MUCH MORE OFTEN COLLECTED BY GIRLS THAN BY BOYS

	808 girls, per cent	808 boys, per cent
(*a*) Objects possessing æsthetic value or appeal.		
Ferns	4	2
Flowers, pressed	11	4
Flowers, paper	7	2
Grass (kinds of)	1.3	0.6
Leaves (kinds of)	14	4
Monograms	1.4	0.5
Moss (kinds of)	3	1
Paper, colored	8	3
Paper, tissue	4	1
Pictures	20	8
Poems (collected)	15	6
Rose petals	4	2
Rocks (kinds of)	9	5
Samples (perfume, etc.)	14	6
Pieces of calico	3	1
Handkerchiefs	15	5
Sachet bags	9	0
(*b*) Objects for personal adornment		
Beads	12	3
Bracelets	7	1
Breastpins	3	1
Ear rings	7	1
Jewelry	13	4
Lace	9	1
Hat pins	1	0
Scarf pins	1	0
Stick pins	1	0
Rings	10	3
(*c*) Objects which possess sentimental value		
Autographs	17	5
Autograph sentiments	14	3
Song books (kinds of)	9	3
Cards (election, merit)	9	4.7
Letters	14	4
Newspaper scraps	7	3
Programs (dance, etc.)	8	3
Souvenirs	12	7
Valentines	28	11
Charm strings	3	1
Clover, four-leaf	13	7
(*d*) Dolls and doll paraphernalia		
Dolls	11	1
Dolls of paper	14	1
Doll buggies	3	0
Doll clothes	12	0
Doll dishes	8	0
Doll hats	6	0
Doll quilts	8	0
(*e*) Household accoutrements		
China (painted)	4	2
Dishes (broken)	4	1
Napkins	7	2
Pieces of quilts	4	1
Spoons	4	1
Thread	4	1
Strings	4	2
Calendars	7	4
Shoes (old ones)	2	1
(*f*) Souvenirs from the schoolroom		
School books	12	6
Compositions	4	1
Drawings	13	7
(*g*) Objects used in playing games		
Jackstones	9	2

TABLE II—ITEMS MUCH MORE OFTEN COLLECTED BY BOYS THAN BY GIRLS

	808 boys, per cent	808 girls, per cent
(*a*) Animal parts and insects		
Beetles	3	1
Snails	2	1
Spiders	2	0
Birds eggs	5	2
Birds beaks and claws	1	0
Birds wings	2	0
Furs	8	2
Gopher skins	1	0
Rabbit ears	7	1
Skeletons	1 1	0 5
Skins	1	0
(*b*) Junk to sell (at least salable)		
Bones	3	1
Bottles	0	4
Brass	5	1
Iron	7	1
Jugs	3	1
Horse shoes	8	3
Lead	8	2
Metals (kinds of)	5	2
Rubber	4	1
Tin	6	1
(*c*) Tobacco souvenirs		
Cigar holders	3	1
Cigar tags	1	1
Cigar tins	1	1
Cigar stamps	7	2
Cigar papers	4	2
Tobacco sacks	5	1
Tobacco tags	2	0
(*d*) Objects associated with war, hunting, fishing, etc		
Arrowheads	7	1
Bullets or cartridges	10	0
Fish hooks	10	2
Flint	7	1
Guns (toy)	5	1
Knives	10	2
Shot	1	0
Wampum	1	0
War relics	5	1
(*e*) Objects used in playing games.		
Marbles	32	5
Marbles (agate).	22	3
Tops	0	2
Kites	7	2
(*f*) Miscellaneous objects		
Badges	6	2
Books, story	0	3
Matches	2	1
Nails	8	2
Neckties	5	1
Oil cans	3	0
Padlocks	5	1
Pant guards,	1	0
Rubber sheets	1	0
Sacks (kinds of)	2	1
Tags, shipping	1	0
Tags, tin	2	0
Trade marks	3	1
Wagon wheels	6	1

Table I (*b*) presents the names of objects which are used by girls for self-adornment. Although most of these articles were collected to a limited degree by boys as well as by girls, they were collected with strikingly greater frequency by girls Furthermore, only one similar object (necktie) appears in the boys' list Noticeable indeed is the fact that Table I (*b*) contains ten items In the conspicuous interest of the girls in this type of collecting and the paucity of interest among boys there is evidence of one clearly defined sex difference in behavior trend

It is apparent readily that this attempt at classification is arbitrary and therefore somewhat inaccurate. For example, in Table I (*c*) are listed the names of objects which possess sentimental value. It is obvious that almost any cherished possession may have a sentimental value It is equally true that practically any of the items listed in Table I may have been a gift or a token, and therefore may have acquired sentimental value. Nevertheless, if one inspects carefully the two lists, he will grant undoubtedly that the girls' list contains many more objects than the boys' which generally arouse affective reaction and sentiment.*

Table I (*d*) sets forth certain items that are collected almost exclusively by girls, namely, dolls, doll clothing, etc This sex difference would scarcely be considered phenomenal This list however considered with list I (*c*) brings to light conspicuous and apparently general interests of girls

From Table I (*f*), it will be noted that the girls collect souvenirs of the schoolroom much more commonly than do boys. No items of a similar nature appear in the boys' list. (See Table II.)

In the several lists presented in Table I (*a*) to (*f*), striking behavior patterns of girls appear. Girls are attracted by objects which may be used for personal adornment; they assemble and appear to cherish much more than boys objects which call forth affection and sentiment; and they collect and keep relics and portable objects which have been associated with school life The latter tendency may be additional evidence of the girls' rather intense liking for school. This liking is less frequently found in the normal boy *

---

* It is of interest that the only items of the entire one hundred ninety that suggest superstitious belief, namely, charm strings, and four-leaf clovers, were reported more often by the girls than by the boys Sex differences in this regard have been reported by the writers in a previous article (See *Journal of Abnormal and Social Psychology*, Vol XXIII, 1928, pp 356–368)

The objects collected by the boys suggest their out door activity and rather vigorous play life This is particularly striking in the first five classifications of Table II Perusal of this table shows that very few of the objects may be described as ones which possess æsthetic or sentimental appeal. And the lists contain few objects which could be used for personal adornment Few of the girls collected objects which, by any breadth of imagination, could be made to fit into the first four classifications of Table II. The things collected by the girls reflect the relatively narrow geographical radius which characterizes their lives The boys' lists, however, reveal their relatively great participation in activities which require a wide geographical sphere and unrestricted outdoor activity.

The rather striking sex differences found in Tables I and II seem to the writers to explain partially a finding reported by Miss Burk more than thirty years ago. Miss Burk reported:

> The boys exceed the girls somewhat in finding and hunting, and considerably in trading and buying The girls exceed the boys very greatly as passive recipients of outside assistance, in having their things given to them by brothers, sisters, parents, uncles, aunts and friends. But this excess of passivity on their part is not balanced by any special decrease in the method of finding, but it rather balances the excess of trading, buying, and winning among the boys[2] (p. 194)

The preceding comment is corroborated by Miss Whitley.

> The percentage of girls depending on gifts to add to their collections is higher at all ages than it is for boys, and consistently lower for trading[3] (p 259)

The items listed in Tables I and II are not easily classifiable according to the "reasons" which prompted the children to collect them. Nevertheless, careful examination of the lists enables the reader to identify certain of the reasons for the sex difference in *methods* employed in making collections In the girls' list the objects which possess sentimental value (See Table I (c)) are ones which possess also little or no practical utility. This statement is true regarding most of the items in the other lists, I (*a*) to (*f*) Many of the items possessing æsthetic and sentimental appeal are so perishable that they have little or no value for trading or selling. Girls appear to obtain many of their collections as gifts and boys actively assemble collections in order that sale or trade may ensue.

---

* This also is indicative of a sex difference in attitude that the writers have previously discussed (See *Education*, Vol. XLIX, 1929, pp. 449–458)

Of course it is true that many objects collected by the boys are of little actual value or utility. But there is this striking difference between the girls' and the boys' lists. All ten of the items listed in Table II (*b*) have a market value* and many of them have little value other than this The girls' objects of personal adornment probably have some market value, but this value is no doubt usually outweighed by the affective appeal of the articles.

There is one other aspect of this point which is worthy of mention. It will be noted that the writers found it necessary to include a "miscellaneous" grouping for the items collected chiefly by boys. The girls' collections were more easily classified than were the boys' This was due doubtless in part to the greater variety of objects collected by the boys. The great variability of this type of behavior among boys is due in part to their relatively great freedom to take part in outdoor life (as compared with girls) and their relatively vigorous, unrestricted play life.

Because they have pockets in which they can carry about more easily than girls certain of the miscellaneous objects collected by them, the boys are clearly in a better position than girls to satisfy whatever desire they may have to trade objects which they have collected. This desire is also more easily met by the boys because of the fact that in their spontaneous play life they traverse a wider geographical area than do girls. This fact has been previously commented upon by the writers[4] (p. 93).

Although the above attempts to explain the sex differences in "method of and reason for collecting" are of interest, it is the judgment of the writers that such theorizing has less value than knowledge of the actually observed tendencies of children If one accept the view that education is a matter of experience, that the curriculum itself is a series of experiences, it becomes at once evident that the boys and girls herein studied are not receiving the same sort of education. The pupils have themselves assisted in making their curricular differentiation, perhaps on the basis of genuine and deep interests. In any event, before the educator will be in a position to direct most effectively the child's education, he should acquaint himself with the child's actual experiences

---

* This is to be interpreted in terms of the juvenile nature of boys' trading and selling

REFERENCES

1. Witty, P A and Lehman, H C.: Further Studies of Children's Interest in Collecting. *Journal of Educational Psychology*, Vol XXI, 1930, pp 112–127.
2. Burk, C. F : The Collecting Instinct. *Pedagogical Seminary*, Vol. VII, 1900, pp 179 to 207 refer to C F. Burk's article
3. Whitley, M T Children's Interest in Collecting. *Journal of Educational Psychology*, Vol. XX, 1929, pp 249–261
4. Lehman, H C and Witty, P. A.. "The Psychology of Play Activities" New York: A S. Barnes and Co., 1927," pp XVIII + 242.

ERRORS, DIFFICULTY, RESOURCEFULNESS, AND SPEED IN THE LEARNING OF BRIGHT AND DULL CHILDREN

FRANK T WILSON

State Teachers College, Buffalo

In a previous study[1] the learning of bright and dull children was compared in a task of learning how to win a game in which two players alternately draw either one or two from a given number of pieces with the object of winning the last piece. Success requires the application of the principle that one's opponent must be forced to draw from a multiple of three at each of his plays It appeared in that study that for groups of children nine and twelve years of age, half of each group at a level of 70 to 80 IQ and half at 110 to 120, the average difference of 30 IQ points gave advantage as measured by amount of work, to the brighter children, and that the average difference of three years favored the older groups The difference in IQ seemed to be a little more important than that in chronological age since, as measured in the study, the bright nine eventually somewhat surpassed the accomplishment of the dull twelve who averaged practically the same mental age.

I. Explanation of the Study

This report gives the result of further study of the groups in this task in regard to the "Lost moves," or errors made. Data for the groups are given in Table I, six additional cases having been included in the later study

Table I —Number of Cases

	Nine years		Twelve years		
	Dull	Bright	Dull	Bright	Total
Boys	8	8	11	7	34
Girls	7	10	7	8	32
Total	15	18	18	15	66

---

[1] Wilson, Frank Thompson "Learning of Bright and Dull Children" Bureau of Publications, Teachers College, N Y.

The following explanation will make the references to the procedure of the task intelligible to the reader. Complete explanations will be found on pages 12–15 of the first study

"Paper clips were used for the pieces. Each subject had thirty trials each day. A trial was considered to be the series of 'draws involved in the reduction of any initial number (of clips) to 0, regardless of whether *S* (the subject) wins or loses.' Peterson, J. C.: The Higher Mental Processes in Learning. *Psychological Monographs*, Vol. XXVIII, No. 7, 1920, p 3."[1]

The learning was carried on for five successive days. The game was explained and illustrated to the subjects with three clips. The task was begun with four clips and each subject, after having won once with four clips, was told that he must win three times in succession, the rule which constituted the criterion of successful learning of each step. The game continued with the four clips. Having won three times in succession with four the task continued in the same manner with 5, 7, 8, 10, 11, 13, 14, 16, and 17 clips, or as far as the subject worked in the one hundred fifty trials of the series. The experimenter was the opponent of the subject, and his draws were always so made that the subject would win if drawing correctly but lose if drawing incorrectly That is to say, the subject could always win if he applied the principle. As far as practicable all other factors in the experiment were kept constant so that the records show in the main, it is believed, differences due to the differences in the subjects of the four groups.

II. TOTAL LOST MOVES

TABLE II.—AVERAGE NUMBER OF LOSING MOVES

	Nine years		Twelve years	
	Dull	Bright	Dull	Bright
Boys	91.3	81	82	59
Girls	87	84	79	79
Total	89.3	82.7	80.9	69.7

Table II gives the average number of losing moves for each group and for boys and girls. There seems little question but that the bright twelve year old group has the best record and the dull nine the poorest. The slight numerical difference between dull twelve and bright nine gives no grounds for deducing a significant real difference

---

[1] *Ibid.*, P. 12.

Differences between groups by sexes as shown in these data are uncertain because of the small number of cases for the sexes, but the comparisons are given as they indicate the desirability of further investigation. The most striking comparison is that of the bright and dull twelve year old girls, who made practically the same showing. The dull nine year old girls did nearly as well as the bright nine year old girls. In the case of the boys the differences are decided between the bright and dull groups at each age, while the two groups of the same mental age score the same. Comparing sexes it seems that for the bright twelve group the boys did very much better than the girls. In the other three groups the differences are probably so small as to be statistically valueless.

Conclusion 1.—From the straight count of losing moves made it seems that success in the task is more probable for the bright older subjects and that mental age is the most significant factor.

III SAME LOST MOVES

There was made next an investigation of successive same losing moves. Does the presence of differences in IQ and real age produce records for successive same losing moves which compare groups otherwise than do the records of total errors? Table III gives the data for the groups by sexes.

TABLE IIIA —AVERAGE NUMBER OF TIMES SAME LOSING MOVES WERE MADE TWO OR MORE TIMES IN SUCCESSION

	Nine years		Twelve years	
	Dull	Bright	Dull	Bright
Boys	14.8	7.5	10.3	6.3
Girls	12.9	10.7	9.9	10.8
Total	13.9	9.3	10.1	8.7

The figures should be read as follows. The average dull nine year old boy made the same losing move twice or oftener in succession, 14.8

times in his one hundred fifty draws, the average bright nine year old boy 7 5 times; etc.

Something new seems to be present in this table The bright nine year old boys score better than the dull twelve year old and the difference is large. In fact the bright nine are nearer the bright twelve than they are to the dull twelve. The girls, again, however, make records nearly the same, except the dull nine who are much poorer than the other girls. Their superiority over dull nine boys is not sufficiently certain to more than note.

The significance of making the same losing move successively is not wholly clear Questioning of subjects after experimentation brought out the information that sometimes the same moves were made in order to study them more carefully to discover other possibilities. Such a process may have been more frequent with the bright Observation and introspection suggest that sometimes the same moves were made because the previous trial was not remembered correctly and the subject thought he was playing differently the second time. Such processes may have been more common with the dull. Table IIIB, although covering very few cases, emphasizes the weakness of the dull nine in repeating the same losing moves many times.

TABLE IIIB —AVERAGE NUMBER OF TIMES SAME LOSING MOVES WERE MADE IN 3 TO 9 SUCCESSIVE TRIALS

	Dull 9	Bright 9	Dull 12	Bright 12
Same moves·				
Three times	1 5	1 4	9	1 0
Four times	1.1	1	.3	3
Five times	3	0	0	0
Eight times	.1			
Nine times	.1			

A complication in interpreting all the data of this study exists in the nature of the differences in the difficulty of the successive steps of the task This is affected by the practice which goes on with learning each step as well as by the number of clips used in the various steps. The subjects had varying amounts of practice on the several steps depending upon how quickly each won three times in succession. Some subjects had much practice on the early steps because of failure to win the three times Others had little practice in the early steps because they did win three times with only few losses Yet the latter apparently doing better with the problem at first often found later steps just as difficult as those who had more trouble at first. For the bright, however, it should be noted that, as they progressed farther than the dull of the same age, they had to work with steps using more chips and, accordingly, having more possibilities in the way of moves With those larger numbers of chips it seems that subjects intentionally repeated same moves in order to study the situation more thoroughly.

Conclusion 2.—Analysis of successive same lost moves seems to indicate that brightness acts to reduce such errors

IV. DIFFICULTY OF STEPS

The difficulty of the steps, which it has been noted just above, is uncertain, has been studied from the data regarding the number of lost moves by steps and groups. Tables IVA, IVB, IVC, and IVD show these data.

TABLE IVA —DIFFICULTY OF STEPS BY GROUPS MEASURED BY THE AVERAGE NUMBER OF ALL LOSING MOVES

Step	4	5	7	8	10	11	13	14	16	17
Dull 9	2 0	2 2	32 7	10 7	31 3	5	1 6	8	.7	
Bright 9	.0	1 2	6 3	12 5	30 1	16.4	4 3	1.3	8	
Dull 12.	7	1 4	7 0	15 8	38 4	11 6	3 3	.9	1 3	05
Bright 12	4	6	6 4	8 5	27 8	11 9	8 4	2 6	2 7	.4

TABLE IVB —DIFFICULTY OF STEPS BY GROUPS MEASURED BY THE AVERAGE NUMBER OF SAME SUCCESSIVE LOSING MOVES

Step	4	5	7	8	10	11	13	14	16	17
Dull 9	40	33	5 80	2 40	4 33	06	.26	00	20	
Bright 9	11	.00	00	1 56	5 27	1 22	22	11	06	
Dull 12	11	11	1 00	2 66	4 39	1.26	50	05	00	
Bright 12.	.07	07	1 00	1.27	4 07	1 27	.73	07	13	

Table IVD is an hypothetical record made up by converting the number of actual lost moves recorded for each group into what those figures hypothetically become if one hundred per cent of each group had been allowed enough trials to solve each step. The arithmetical sources for Table IVD are found in Tables IVA and IVC. There is considerable probability that if all subjects had been permitted to work until each step had been solved as far as shown in Table IVD the actual records would be different as the practice obtained at one step would affect the work on the following steps. But even granting such probability Table IVD doubtless represents the facts in regard to comparative difficulty of steps more truly than either of the raw data in Tables IVA and IVB. If objection is made to this assumption, however, the raw figures may be referred to as the same relative differences and comparisons of groups are found in them, although not such striking ones The figures for the larger numbered steps have little significance because few subjects reached them It seems, then that attention is warranted chiefly to the first part of the series.

TABLE IVC.—PERCENTAGE OF EACH GROUP SOLVING THE VARIOUS STEPS OF THE TASK

Step	5	7	8	10	11	13	14	16	17
Dull 9 . . .	100	80	67	7	7	7	7		
Bright 9 . ..	100	100	100	83	44	17	7		
Dull 12 .	100	100	100	67	28	11	11	7	
Bright 12 . ..	100	100	100	80	73	40	27	20	13

TABLE IVD—DIFFICULTY OF STEPS BY GROUPS MEASURED BY THE AVERAGE NUMBER OF ALL LOSING MOVES CORRECTED FOR ONE HUNDRED PER CENT SOLVING EACH STEP

Step	4	5	7	8	10	11	13	14
Dull 9 . .	2.0	2.2	40.9	25.0	447			
Bright 9	6	1.2	6.3	12.5	47.1	37.0	25.3	
Dull 12.	.7	1.4	7.6	15.8	57.3	41.4	30	
Bright 12	.4	6	6.4	8.5	34.8	16.3	21.0	9.6

The dull nine group found steps four and five much more difficult than did the other groups Step seven became a much more difficult problem for all groups than the preceding ones, but for the dull nine it appears to have been very much more difficult, only eighty per cent of the group ultimately solving it and the average number of lost moves at that step being 33, or corrected, 41. Step eight appears to have been easier for the dull nine, at least after the practice they had had on the previous steps. There is a slight drop in the percentage which solved this step, but even so the corrected number of lost moves shows a smaller figure for step eight than for step seven For the other groups step eight seems to be harder than seven. The next step jumps the losing moves up to very much higher figures, and from the table of the percentage who solve step ten this one is Waterloo to the dull nine. Only seven per cent survive its demands. The corrected average number of lost moves thereby becomes 447. There seems to be no doubt but that the difficulty of step ten, even after the previous practice of the task, is very severe for all groups. The next step, eleven, is not quite so hard for the remaining three groups, although it eliminates more than another one-third of the dull twelve For the three groups it is harder than any previous step, excepting step ten Possibly the succeeding steps are easier. The small amount of data for them does not warrant conclusions and none are therefore made.

Conclusion 3 —Apparently all groups found step ten the hardest as it was met with in this series. Relatively it was very much the hardest for the dull nine. For all groups but dull nine, the preceding steps increased in difficulty and the succeeding steps decreased in difficulty For dull nine step seven presented an extremely acute difficulty and step eight considerably less difficulty than either step seven or step ten.

V. Resourcefulness

Whether or not this term is the correct one to use, the data given in Table V show in a comparative way to what extent the different groups used successive same losing moves and other losing moves The figures for each group give the per cent that the moves at each step were of the total number of the whole series of steps.

The table should be read as follows· The dull nine made three per cent of its successive same losing moves at step four; two per cent at step five; etc It would seem that, if the proportion of successive same losing moves was greater than that of the other losing moves at the beginning of the task and if the proportion decreased toward the latter end, it would be reasonable to hold that subjects were abandoning a method of little or no return for some other method of more promise.

The figures in the table indicate a slight superiority of the bright groups over the dull, especially if the per cents are totaled for steps four to ten Bright twelve, which made the best progress in solving the problem, has a much better record than dull twelve and the bright nine show a better score than dull nine.

TABLE V —PER CENT AT EACH STEP THAT SAME LOSING MOVES ARE OF TOTAL SAME LOSING MOVES COMPARED WITH PER CENT AT EACH STEP THAT ALL OTHER LOSING MOVES ARE OF TOTAL ALL OTHER LOSING MOVES

	Steps							Total steps 4–10
	4	5	7	8	10	11	13	
Dull 9								
Same losing moves	3	2	43	18	32	.		98
Other losing moves	3	3	37	20	37	..		100
Bright 9								
Same losing moves	1	1	7	17	57	13	.	83
Other losing moves	1	2	8	15	47	21	6	73
Dull 12								
Same losing moves	1	1	10	27	44	13	5	83
Other losing moves	1	2	10	20	50	14	4	83
Bright 12								
Same losing moves	1	1	12	15	47	15	8	76
Other losing moves	1	1	9	12	40	18	13	63

Conclusion 4.—In the kind of "resourcefulness" shown in Table V brightness seems to be significant

VI. SPEED OF LEARNING

TABLE VIA.—AVERAGE NUMBER OF TOTAL LOSING MOVES FOR EACH DAY

Day	Boys		Girls		Total	
	Dull	Bright	Dull	Bright	Dull	Bright
	Nine Years					
1	17.3	11.3	12.0	9.2	29.3	20.5
2	15.0	14.4	17.3	16.4	32.3	30.8
3	17.8	12.6	15.6	19.2	33.4	31.8
4	22.7	24.3	20.7	20.3	43.3	44.6
5	20.0	18.1	21.3	19.2	41.3	37.3
Total .	91.3	80.6	87.0	84.3	179.6	165.0
	Twelve Years					
1	10.3	7.4	9.3	8.8	19.6	16.2
2	16.6	11.1	15.0	15.3	31.6	26.4
3	18.7	14.9	18.1	17.9	36.8	32.8
4	16.5	13.1	17.7	17.5	34.2	30.6
5	19.8	12.1	19.3	19.9	39.1	32.0
Total . . .	81.9	58.7	79.4	79.3	161.3	138.0

TABLE VIB.—AVERAGE NUMBER OF SAME LOSING MOVES EACH DAY

Day	Boys		Girls		Total	
	Dull	Bright	Dull	Bright	Dull	Bright
	Nine Years					
1	2.4	1.1	3.6	1.3	2.9	1.2
2	2.3	1.5	2.8	1.6	2.5	1.6
3	1.4	2.9	2.9	1.1	2.2	2.1
4	3.3	2.8	2.6	2.1	2.9	2.5
5	3.4	2.4	3.3	1.4	3.3	1.9
Total . .	12.9	10.7	14.8	7.5	13.8	9.3
	Twelve Years					
1	.9	1.1	1.1	1.1	1.0	1.1
2	2.4	1.8	2.3	1.1	2.3	1.5
3	1.7	3.2	2.1	1.3	1.0	2.3
4	1.4	2.5	2.3	1.6	2.0	2.1
5	3.3	2.3	2.5	1.1	2.8	1.7
Total.	9.7	10.9	10.3	6.2	10.0	8.7

Table VIA gives the figures for the average number of total losing moves by sexes and groups for each daily set of thirty trials Table VIB gives the figures for the average number of same losing moves by sexes and groups for each daily set of thirty trials This arrangement of the data shows something as to the rate at which errors decreased or increased during the series of one hundred fifty trials Using the bright twelve year old boys as suggestive of the most efficient learning it is found that in both tables that group increases its average number of errors through the second and third days and then progressively decreases the number for the fourth and fifth days. With more or less irregularity either the opposite takes place with all the other groups, that is each tends to increase its errors from day to day, or they fail to materially reduce the errors The bright nine year old boys offer the most irregularity. One is in doubt in regard to this group whether to believe the number of errors tends to increase throughout the first four days and then tends to decrease, or whether the apparent great difficulty of the fourth day really belongs to the third and fourth days and means that progress remains about the same after the second day.

Conclusion 5—Measured by frequency of errors it seems that brightness operates to decrease errors more quickly.

SUMMARY DISCUSSION

Comparison of bright and dull children in certain noted reactions of the complex process of the task of this study shows for the groups of the experiment that on the whole the older and brighter children react more efficiently than the duller and younger. The noted differences are not, however, certain differences in all cases The certainty seems best for bright twelve year old boys

Two observations may be made in view of these findings Each has merit and the two lead to an hypothesis of importance. First, whatever real differences may exist between the four groups studied, each group and every last subject in each group, made appreciable progress along the same lines of effort. Dull and bright, young and old solved steps in the problem as measured by a practical, common-sense criterion of every day kind, namely: Repeated success in getting the desired result—in this case winning the last chip three times in succession at each step It took the dull and younger a little longer to advance from step to step, but by the same measure of success they did advance several steps They made errors, to be sure, but the brighter and older ones also made errors and the same kind of errors Individually some of the brighter ones, also, made more errors and progressed more slowly than some of the dull ones

This observation of the general ability of the dull groups to progress along the same lines as the brighter ones is an interesting one *to keep in mind in view of the often heard statement that dull folks* are not headed in the same direction as bright ones.

The second observation offered is that perhaps the gross nature of the measurements used in the records failed to reveal the importance of the differences really present Three illustrations of such apparently small but significant differences, are found in reputable mental tests. Gesell in his schedules for studying the mental growth of the preschool child rates an infant who "unmistakably articulates" two words in addition to ma-ma or da-da as a very high nine months old child; while five words, only three more, places him as very high at twelve months. That means that at this period of life just three more words make a warrantable addition of 33⅓ per cent in estimated mental development In the Stanford revision of the Binet test the exact repetition of 12-13 syllables is given for the fourth year; the repetition of 16-18 syllables for the sixth year; 20-22 syllables for the tenth, 28 syllables for the sixteenth Thirty words in the vocabulary test place one at ten years; forty at twelve; fifty at fourteen, sixty-five at sixteen, and seventy-five at eighteen. The nature of the task studied in this report may very well be so complex and demanding that the small differences recorded in the data really mean great differences in ability

If this conclusion is tenable a following general conclusion is justified and, it seems, highly significant. It is, in words of the day, that the ability of the dull and young subjects in this complicated problem is not to be laughed at How far they have gone from the true zero point can not be said, but if the small recorded difference between their ability and that of the bright is so much that it stands for a large real difference, then the differences between the arithmetical zero of these data and the points scored by the dull groups must stand for a truly great amount of ability, certainly for so much that in it is promise, not despair

Final Conclusion.—Dull nine and twelve year old boys and girls found the complex task of learning to win a game, success in which depended upon the application of an underlying principle, more difficult than did bright nine and twelve year old boys and girls, when their success was estimated in terms of frequency and repetition of errors. The dull did progress, however, and the bright children made the same kind of errors but fewer of them.

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII April, 1931 Number 4

SELF-CULTIVATION AND THE CREATIVE ACT: ISSUES AND CRITERIA[1]

HAROLD RUGG

Teachers College, Columbia University

1

No contribution of the child-centered schools is greater than the discovery of the principle that only an artist-teacher can discover and develop the artist in the child. Only one who has lived through the art experience can provide art experiences for the children; that is, one must himself have the attitude of a creative person in order to develop children into creative persons. This is the important generalization that emerges from the scores of creative artists in our new *schools during the past fifteen years.*

But the cultured person is both pragmatist and artist, both a maker and doer and a creative appreciator of life The complete educational program, therefore, must embrace the attitudes, concepts and techniques of the creative and appreciative individual as well as those of the problem-solver Now to develop the instrumental values and activities of life a monumental library has already been produced under the leadership of Mr Dewey. But of the criteria for the education of the Man-as-Artist little has been said.

It was to the concepts of the artist that we turned in our attempt to sketch a balanced portrait of the integral person. In him we perceived the guiding concept: The integrity of one's own self . . . one's authentic inner truth as the true criterion for judgment and for conduct . . . admiration for every well-thought-out personal philosophy It is the Man-as-Artist who is sensitive to the criterion of integrity, who is dominated by the attitude of appreciative awareness The true craftsman is he who stresses Feeling-Import, who gives creative desire a place coordinate with intelligence He sees man whole and in turn visualizes the nation as a multitude of integral persons. Corresponding, it is the artist in the school who is concerned with the production of persons, not primarily with developing professional poets, painters, actors, thinkers, or musicians. Hence the source for a psychology of the creative act lies in the experience and the vision of the artist himself

---

[1] Chapter XIX of the author's forthcoming book, "Culture and Education in America." Harcourt, Brace and Company, New York (In press)

In addition to possessing a sensitiveness to the norms and criteria of the creative attitude there is another fundamental qualification for serving as the artist-teacher That is, the sensitiveness to the potential artist in the child and his methods of work The development of creative education in the schools depends more crucially upon the adoption of the "drawing-out" attitude by the teacher than it does upon his mastery of technical knowledge and skill in manipulating the materials of one art. The creative teacher has this attitude He regards every child as a potential artist, each in his own medium of creative power. Hence, the artist is a sensitive listener to childhood As Miss Levin puts it: "You have to feel the thing the child wants to do, to think his thought, in short, become a child yourself"

The educational conclusion from such postulates is clear In order that children shall become self-expressive craftsmen with words or tone, with clay, wood or stone, with light and shade, the teacher must be a craftsman with those materials. Thus, and only thus, can he become sensitive to the potential person in the child This was the epoch-making message of Hughes Mearns. It was likewise the stimulation given by Satis Coleman in her "creative music" So the theme recurs with the materials of painting, sculpture, the dance, or the drama. Through each medium of self-expression children develop the ability to express themselves honestly, creatively, and to grow as persons only to the extent that the teacher's attitude and procedures provide for them.

Thus, the significance of the entrance of the artist into the school is as great as his discovery of the artist in the child It is through creative production in the school, in the home, in every social agency that children receive practice in developing the integral of one's self They learn to measure themselves against critical inner standards Children must ask: Is this poem I have written really "I"? Is this house I have made, this music I have played, utensil I have constructed, brief I have prepared, oration I have delivered, as close an approximation of my true self as I can make it? This is the criterion which the artist ruthlessly applies to himself, and this is what he has contributed to the creative phases of the new education.

He has recognized attitudes, points of view, and techniques which are indispensable in the education of the cultured man Thus he *offers an important contribution to general education as well as to the* development of potential artists.

2

The scientific attitude is analytic, and the method of work concentrates upon the systematic collection of facts measured on scales of equal units, the discovery of recurrences, uniformities, "law"

The artist's attitude, however, is integrating, all-embracing, appreciative. The artist "measures" but against the unique inner *norms of his own peculiar experience, not against external standards.* Correspondingly, his goal is nonuniformity, the unique individual thing Whereas science emphasized the adjustment of the individual to an external norm and seeks the confirmation of generalizations, art on the contrary shuns repetition and denies the possibility of confirmation or refutation

These distinctions set the stage for our study of the difference *between the methods of work involved in three essential kinds* of activity: (1) Intelligent problem-solving, (2) creative production, (3) appreciation Phrased another way, the question amounts to this· What are the likenesses and differences among these processes? Students of the creative act maintain that there is a difference between the process of problem-solving (in which assimilation plays the leading rôle) and that of creative self-expression and contemplative awareness. The instrumentalists deny this They maintain that the assimilative act and the creative act are merely differing aspects of the same general procedure of learning. Always protagonists of the unity of experience, they maintain that those who distinguish between "assimilation" and "creation" are resorting to a dualism Since the discussion throughout the book has emphasized the unified character of human experience, that criticism does not apply in the present instance. The distinctions which are now to be pointed out are of another type

At this point we must remind ourselves of the current tendency to apply the word "creative" to any kind of active vigorous learning. Such a tendency is conspicuously evident just now among educationists Lectures by professors of education and others subsume under the caption "creative" the most obvious kinds of repetitive learning, mastery of skills, and acquiring of information Nothing but confusion can come from such a careless use of meanings and vocabulary

Next, a word concerning the data and the method of my analysis. The data are the subjective materials of experience, and the method is that of introspection, or rather retrospection We are studying the mental and emotional experience undergone in the *creative* act. It is, therefore, only by the introspection of the *creative* artist that the experiential data of the process can be explored. No person who has not experienced this process can generalize concerning it, and no objective measure of products can lay bare the process itself

Hence my description of the creative process is based upon occasional flashes obtained in the autobiographical literature of creative artists and creative scientific men, from similar examples of intuitive understanding obtained in conversations with such persons and from my own introspections. (In earlier years I had prolonged contact with the processes of the scientific method both in physical and intellectual technology In recent years I have devoted considerable energy to the creative arts)

The techniques of introspection and retrospection must not be despised They are the only possible means of exploring "processes" of learning. They are indeed the techniques used by Professor Dewey in his famous analysis of "the complete act of thought"[1] (that is, the complete act of problem-solving, for Dewey is considering only one type of thought) The validity of that subjective analysis was established, of course, by the confirmation that came from the cumulative, introspective analyses of other students of problem-solving. The validity of my present analysis of the creative process must be established in the same way I am only too well aware of the difficulty which the creative artist encounters in giving utterance to the kaleidoscopic succession of his complex inner states. Nevertheless we must attempt the task and to do so we must put ourselves in the attitude of the artist Only so can we make a valid record of our experiences

Furthermore, as I said before, no objective measurement of products or of overt behavior will portray the inner process itself That is clearly illustrated in the monumental work of Thorndike and

---

[1] Dewey, John "How We Think." D C Heath and Company, New York, 1913, Chap. V

Judd in reading, and of Freeman in handwriting Thorndike measured the product of reading; Judd and Freeman measured the overt signs of the processes themselves, that, is by the photography of eye-movements in reading, by hand and finger movements in writing But even the latter did not record the experiences which produced the behavior of the product What they are can only be inferred and, indeed, only approximated by the careful introspection of the person himself.

3

What, then, is our task? It is the analysis of three ways of knowing. Problem-solving, creating, and appreciating Let us begin with an analysis of the four aspects of the learning process. First, the attitude of orientation; second, the initiation of the act; third, the ongoing process itself, fourth, the definiteness with which the achievement of the goals can be ascertained, that is, the possibility of confirmation or refutation.

In analyzing these we shall postulate the unified nature of experience We shall conceive of each human response as an integration of physiology, intellect, and emotion. No dualistic separation of faculties or elements of response is implied in my discussion. On the contrary, it is assumed that the organism responds as a unit, and every act is conceived as a fusion of sensori-motor set (attitude), of meaning, of generalization, of language, gesture, and other overt movement. Life is a succession of infinitesimal experiences, each a complex weld of the many kinds of process with which the human organism has the capacity to respond The human act is not mosaic; it is fusion.

Nevertheless, our argument is that all human acts are not alike For example, the tendency to flee from a situation is unlike the tendency to embrace it and we give different names to them. The emotion of anger is unlike the emotion of love and their descriptions vary accordingly One act is oriented by one attitude, another by a very different one One makes much use of meaning and generalization, is predominantly intellectual. Another, however, is predominantly motor response, calling into play but few intellectual elements. Still another may be hyper-sensory, involving little overt movement but much internal kinesthetic response—a gathering-together process highly charged with emotion Thus the composition of human acts varies greatly. Hence, in exploring the acts of problem-solving, creating and appreciating we shall expect to find similarities and differences in their constituent elements.

4

Consider, first, the attitudes orienting the act of problem-solving. In confronting a problem, the worker is oriented outward. The conditions of the problem are "given" Here are some examples: One has to find the distance between two points, to discover the combination of elements which make up an unknown; to design an engine of known-in-advance horse power, cylinders, and the like, to determine the most favorable distribution of practice required to produce a desired amount of skill, or to determine the dimensions of an I-beam which will withstand known-in-advance loads.

In each of these "problems" the attitude is set in reference to external needs To grasp the problem, the individual must adopt the attitude necessary to understand the conditions set by it. As Dewey has said, unless the individual can perform an indicated set of operations, he cannot respond with its meaning, that is, unless he adopts the attitude appropriate to the problem he cannot understand it. This is the essence of the active psychology of meaning built up primarily through the efforts of Peirce, James and Dewey It is only by striking the attitude rigorously determined by conditions outside his own experience, external to his background of meaning and generalization, that the problem-solver successfully recognizes the "felt difficulty" in the "forked-road" situation.

In the creative attitude, however, the orientation is inward It is subjective, not objective, as in problem-solving The creating process is propelled by an inner urge to objectify moods, to portray overtly personal integrations of meaning, generalization, and emotion. The drive may be to write a poetic phrase or line or stanza, to portray something with pencil or brush, to put together a new combination of tones or bodily movements that will objectify a fusion of ideas and feeling But the attitude adopted in the initial stage in the creative act is determined by reference to the subjective, inner experience of the individual

There is also a second distinction. Whereas the "problem" of the problem-solver is external to the individual, the "problem" of the artist is internal. There is a difference in definiteness. Problem-solving is focussed with sharpness upon conditions prescribed in the external world, such as the loads, span, strength of materials, etc involved in the design of the I-beam, the location of points involved in the determination of distance, or the prescribed horse power, number of cylinders, etc in the design of the engine The problem-solver must adjust with exactitude to these externally prescribed requirements

Not so with the orienting attitude in the creative act It consists at first of little more than a vague restlessness, an undefined desire to express in an external product the internal experience of the individual This gives us, indeed, an important cue to the difference between problem-solving and creating; that is, the unchanging rigor and clarity of definition of the externally-set problem and the constantly changing indefinite character of the artist's subjective vision.

Now compare the appreciative attitude with that of problem-solving and creating The first has resemblance to both the second and third, but it is much more like the creative attitude than that of problem-solving. What are its elements? Although it is stimulated by external conditions, it is not really oriented externally. It is oriented by the internal personal gathering-together of the self. In appreciating a tone-poem, a dance, a painting, statue or a building, the individual does not strive to comprehend the meaning of the artist, the dancer, or the designer He strives only to catch the *coordinate whatever reverberations come from the observed thing* Although stimulated from the outside he makes the response, the interpretation which his own mood, experience, and needs call forth He "appreciates" in terms of what he is, what he feels and understands It is a sort of confident gathering together of the whole personality It is an attitude of awareness, as well as one of critique; it is all-embracing rather than analytic

In all three of these attitudes there is of course meaning, generalization, physical adjustment and emotional content, but there are distinctive differences in their amount and integration. The appreciative and creative attitudes are effective only to the degree to which they are highly charged with emotion The problem-solving attitude is effective only to the degree that the worker maintains emotion at a low ebb. This does not mean that the problem-solver is not also gathered-together emotionally. He is concentrated intently upon his task Sometimes a thrilling orientation is perhaps essential to success On the other hand it is frequently the emotional intensity of concentration which inhibits the making of appropriate generalizations

So much, then, for the differences in orientation of problem-solving, creating and appreciating.

5

Consider, next the similarities and differences in launching the acts themselves. Dewey, in his classic analysis of the complete act of thought, leads us to the step immediately following the recognition of the problem—the flashing up of suggestions or solutions In the case of problem-solving these solutions are hypotheses drawn from known data They are tested against the requirements of the problem. They are generalizations drawn from "facts" and are accepted only when under scrutiny they fit the facts

In the creative act also this flash-like succession of fused emotion, meaning and movement resembles closely the successive appearance of solutions in problem-solving. But, as in the original orientation of the work, there are sharp distinctions between the two processes The "suggestions" which flash up to the problem-solver are hypotheses from known data. The meaning of each is precisely fixed; generalization must fit the conditions set by the problem itself. In the creative process, however, the suggestions for modifying the art product are measured against the changing subjective experience of the person himself. "Solutions" are accepted only as he perceives that they correspond to his felt moods Thus the point of reference in this creative enterprise is subjective.

In appreciating, "analysis" of a kind undoubtedly does play a part as it does in problem-solving and in creating. In listening to a symphony our enjoyment is enhanced, for example, by noting the recurrence of themes, the manner in which the tones of particular groups of instruments merge with others, and the like. Our appreciation of a poem or of prose writing undoubtedly is augmented by mastery of form and style which opens to our emotional comprehension varied channels of enjoyment. Illustrations could be multiplied for other media of expression. There is no doubt, therefore, that in the fullest development of the appreciative act, the mind attends to separate phases or aspects of the process as well as the ensemble

As in the creative act, however, there is a fundamental distinction as to point of reference There is no attempt to make the symphony, the landscape, the statue fit into prescribed objective standards. We take from the situation what we can. We integrate it with our on-flow of experience, but we must not interrupt the process of receptive awareness to dissect critically and judge the validity of the art product If we should do the latter, then the process becomes one of problem solving and not of appreciation.

6

Third, consider the "methods of work," the on-going procedure itself in problem-solving, creating and appreciating

In problem-solving the individual "collects," classifies and compares facts If they are multitudinous, he condenses and summarizes them by statistical devices These facts are measured against the scales of approximately equal units. Since I have discussed this process in earlier chapters, I shall merely summarize the argument by saying that the whole process of problem-solving is one of drawing generalizations which fit the external requirements of the problem.

How different is the creative method of work. The "facts" are the lines, words, or phrases that well up out of the inner recesses of the artist's experience. They are the tentative modelings in which the sensitive hands of the sculptor or painter attempts to objectify his moods and his vision They are the imagined arrangements of light and shade, color and materials in the stage-set and costuming of the scenic designer They are the apparently uncoordinated arrangements of materials, dimensions, machine or instrument parts of the mechanical inventor. They are the interpretive bodily movements of the dancer responding to music.

These are the "facts" of the creative process Are they measured against external and precisely standardized norms? No, they are measured against the artist's unique inner sense of the total situation to which he is responding. The succession of experiences is a procession of changing, complex, unique situations. Hence, the utter impossibility that the artist shall systematically assemble meanings, lines, words, color, what-not in terms of externally known-in-advance conditions As with the orientation and initiation of the creative act, the on-going process of the work itself is essentially objectification of moods, slowly, defining images, gradually shaping mental and emotional complexes

Obviously, meaning and generalization play a selective and interpretative rôle as well as emotion Again we say that the creative experience is a fusion of physiological processes, meaning, kinaesthetic and emotional reaction. Hence, intellectual elements play a part—

perhaps a guiding, orienting part. Definitely what part meaning-reacted-through-words plays in determining the actual step-by-step advance of the artist's work we cannot say at the present stage of the analysis of the creative act We lack definitive retrospective studies by creative artists of the development of their work

It is clear, however, that the "analysis" that leads the artist to "correct" words or phrases, shapes, lines, arrangements of material, light and shade, combinations of tones, bodily movements is a *fusion* of emotion and intellectual meaning. Ideas play a decidedly important selective rôle in dealing with media that are primarily intellectual, that is, words in poetry and essay, and symbols in mathematical invention But they play a very minor building rôle in dealing with the media of the graphic and plastic arts, music and the dance Here feeling-import ousts ideas as the directive agent. Thus it is clear that *analysis* plays a part in the carrying on of the creative method of work as well as in problem-solving

This shows itself conspicuously in the conscious effort with which the artist gathers himself together, determinedly focuses his mind upon the manifestations of his moods, and drives himself to define the vision which serves as his inner goal Thus he adopts an attitude of ruthless criticism of the objective portrayal which is appearing on his canvas, *in his statue or musical composition* To the best of his ability does it correspond with his inner vision? New relationships are consciously sought, new combinations of words, lines, tones, shapes There is a rigorous search for new elements in design. There is a dogged determination to discover and build upon unifying themes. This process inevitably sharpens the vision as well as produces a maturing objective expression Thus the on-going analytical process of the creative act clarifies both the inner subjective state and the outer objective product.

The goal in the act of appreciation, as we have already pointed out, is fullest awareness The governing attitude is a gathered-together receptiveness, an attempt to embrace all the reverberations that radiate from the words, the tones, the lines, the masses of color, or the movements of the body, in the stimulating situation The process is essentially one of "listening," not of making and doing. Here is an example of "assimilation" that is closely akin to the creative act and quite foreign to the "assimilation" of the problem-solving act. In appreciating we assimilate stimuli of tone, movement, or word-meaning, into our current moods We "enjoy," thrill over the new assimilation.

We build it into our "experience" We accept it, embrace it, are lifted up, carried onward

"Analysis" plays its constructive part in appreciating also, but it is more like creating than problem-solving. In music, for example, the individual "notes" the use of specially related tones and chords, the recurrence of basal themes, the integral unity of the composition To the extent that he is a master of the technical forms of construction, his enjoyment is enhanced by the recognition of the artist's techniques. Thus analysis leads the listener to create a new inner product; his fusion of impressions is unique; it is "himself" living on the maximum heights of appreciation

But he does not analyze in order to reproduce in himself what the artist felt and saw in creating the product [1] If he does so it happens by *chance*, not by design. That is, he does not analyze to "solve an externally-set problem" If he questions, for example: "Do I get the artist's intent?" he at that instant turns from appreciating to problem-solving.

7

There is a fourth difference to be noted—namely, the difference in the definiteness with which the goal of the worker can be visualized and the success or failure of achieving it ascertained The problem-solver knows in advance the problem he is to solve. It is the discovery of the "law" of relationship between factors that change together. His method is to discover the exact combination of factors—materials, volumes, loads, pressures, repetitions, lengths, masses and the like—which will bring about the stated conditions The very statement of the problem fixes precisely the goal of the worker, it is unchanging throughout the entire mental process

Not so in the production of a creative thing The goal is the clarification of the artist's vision and its objectification He must see and feel clearly enough to portray his vision on a canvas, or in a poem or dance or musical production The goal of the creative artist, therefore, changes constantly The nature of the creative process is tentative and hesitant, there are constant interlineations, erasures, the giving-up of old achievements, the adoption of new experimental arrangements Hence, also, there is the continual attitude of discontent.

---

[1] Hence the futility of most "criticism" in the arts!

One more important distinction remains to be noted between the outcome of problem-solving and creative production. The solution of a scientific problem can be definitely confirmed or refuted, that of the artist's problem cannot. It is of the essential nature of scientific work that a generalization once achieved in the solution of a problem is susceptible of exact confirmation by independent workers at other times and places. Indeed, inferences drawn by the scientific worker are not regarded as scientific "law" unless they can be exactly confirmed by other workers utilizing the same set of procedures. Thus, a central element of the scientific method is the inevitability of the discovery of recurrence.

Exactly the opposite is true with the creative product The goal of creative production must be a unique thing It is a painting, a poem, a tone poem, an oration, which is an objective portrait of an inner personality Hence the impossibility of confirmation, or of refutation of the product of such a personality by another

By what standards shall its confirmation be measured? By whom is it to be confirmed or refuted? The product is the artist's personal record of Self. At any given moment no two human beings in the world are even approximately alike Thus it is inconceivable that, except by the remotest operation of chance alone that the peculiar fusion of feeling-import, meaning and bodily understanding achieved by an artist could ever be achieved by another. And if they were so achieved, it would not be a "confirmation" of the original artist's "generalization" It would either be sheer imitation, or a new original product of the second artist. If it were the latter, it must be measured against his personal vision, not against that of the first artist

8

Our comparison of the acts of problem-solving and of creating throws light upon the current confusion of thought concerning "representative art" and "creative art" in the schools We can see now that these differ definitely, but that each is necessary in the education of the cultured youth and in the progressive reconstruction of society. Each plays its important rôle in developing tolerant understanding and dynamic participation in modern life. To get the issue clearly before us let us first illustrate what is meant by "representation," by "representative" art.

From the classrooms of the child-centered schools emerge a host of thrilling illustrations of the use of representative art In a first grade unit on local community life is a model of a miniature city. Grocery store, post office, city hall, town park, railroad station, each item made by a pupil in the class. Together these part-way creative products constitute a representation of the community.

Likewise, in the fourth grade is a more mature representation of the community, a map drawn roughly to scale showing the districts of the city, rivers, chief industries, names of transporting and communicating facilities, schools, municipal government buildings, and residence districts Here, too, is "representation," of a more mature type, however, than that of the first grade "play city."

Consider another—a dramatization of a play written by children in the fifth grade depicting stages of community development. "Research" has been carried on by the young people to make the costumes, appearance of houses, farms, factories and stores "represent" fairly the civilization depicted.

A sixth grade class has been studying the European background of American history Witness the models of medieval castles and towns, a painting depicting life on an English manor, another illustrating transportation in the sixteenth century.

These examples, culled from the experimentation of the new schools throw into clear relief both the importance of representative art in the schools and of its essential characteristics. At the same time they provide us with clear illustration of the distinctions and the similarities between representative art and creative art. Note these succinctly.

First, representation employs the essential attitudes and procedures of problem-solving. The individual confronts a problem, namely that of portraying with relative fidelity the life of the region, group or period, or the structure and form of the plant, animal or society under consideration. The orientation is upon a stated set of conditions, factors, needs, it is outward toward an externally set problem

Second, the student collects "facts" as in problem-solving He searches for historically correct models, for information concerning costumes, modes of transportation, language, what-not In "representing" life he is obligated to approximate the "truth" He attempts to get a "true" feeling for the life of the group, the community, the nation in the period under consideration

In the third place, a necessary measure of one's success in representation is the degree to which he conveys a message to an audience, the degree to which he portrays to others a corresponding feeling for the life which is being depicted This is the very essence of representative art. That it is essential to the sound carrying on of the education of youth is clear Scholarly, technically sound, representative art must occupy a growing sphere of influence in our progressive schools

But imperatively essential as it is, representative art must not be confused with creative art in which the individual's objective portrayal is not controlled by the desire to reproduce either current or earlier forms and conditions of life. The outline which we have already given of the representative process of the creative process show clearly the likeness of the former to problem-solving and the clear distinctions of the latter

Two conclusions of crucial importance are possible as a result of this discussion: *First*, representative art and creative art are two different things; second, both are necessary in our schools Representative art will supply a crucially needed means of artistic expression in building a clear understanding of our changing society Creative art is indispensable to the complete development of the cultured man

TETRAD-DIFFERENCES FOR VERBAL SUBTESTS*

WILLIAM STEPHENSON

University College, London University

INTRODUCTION

A group test† of "verbal" subtests was applied by the author to 1037 girls, followed on the next day (per testing group) by a group test of "non-verbal" subtests. The latter has received attention in a previous paper.[7] The purpose of the present paper is to determine what we can of the factor-characteristics of the verbal subtests, relative to the verbal subtests themselves. Data will be gathered concerning the satisfaction of the Theory of Two Additive Factors by verbal material. Further consideration must needs be given to error other than sampling in the tetrad-differences: As we suggested in the previous article,[7] sampling is likely to be only one of many sources of error. For small populations sampling error is large in comparison with most other sources of error in tetrads that we have knowledge of. But while sampling error is diminished by increase of population, we may have no such control for other errors: Errors insignificant (relative to sampling) for small populations must needs become more and more observable as sampling error is diminished. Thus, contemporaneously with examination of material in terms of the Theory of Two Factors, we must give consideration to error other than sampling, and knowledge of such errors is one of the objects of our work.

THE VERBAL SUBTESTS

The verbal group test consisted of eight subtests, named "verbal" because the test-units made use of printed words, phrases, or sentences or paragraphs. Samples of the subtest test-units follow, each test-unit showing correct responses. Each subtest will be known hereafter by its number 1, 2, etc.

Subtest 1.—Synonyms (inventive) (Responses in italics.)
1 tall high
2 sharp quick
Subtest 2.—Sentence completion
1 Birds build their nests in Spring

---

* The author is greatly indebted to Professor C. Spearman, under whom the work reported here was accomplished.

† *Journal of Educational Psychology*, March, 1931.

Subtest 3.—Classification

1. cap stocking coat (toffee girl wear scarf nut)
2. dog cow elephant (brick pint kitten coal tin)

Subtest 4.—Interchanged words

1. greenhouses are grown in tomatoes
2. The dark had long girl hair.

Subtest 5.—Opposites.

1. come fill go ten
2. every success once failure

Subtest 6.—Analogies.

1. cat: kitten : dog: ______ (horse puppy foal mice catkin)

Subtest 7.—Always has

1. A man always has a (cigar body wife money head)
2. A bird always has (cage seed wings nests legs)

Subtest 8.—Following directions

1. Write the first letter of the alphabet (A)
2. Put an "x" under an "o" in the answer space ($\overset{\text{o}}{\text{x}}$)

Each subtest occupied one side of a sheet of paper, quarto size. The material was typewritten-cyclostyled, print being $\frac{1}{10}$ inch high (small letters). The stapled sheets were given to the testees with subtest No. 1 showing, and the various matters of test routine were gone through with subtest No. 1 as a sample. Each subtest was prefaced by a set of six test-units displayed and worked through on the classroom blackboard, the test-units being printed and spaced as in the particular subtest.

TABLE I

Subtest	Number of test-units	Time allowed in minutes	Mean of crude scores	Sigma of crude scores
1. Synonyms	20	2½	8 85	3 09
2. Sentence completion	25	4	9 98	3 50
3. Classification	24	2½	7 83	3 22
4. Interchanged words	20	3	5 35	2 33
5. Opposites	26	3	8 32	3 18
6. Analogies	25	2½	7 55	4 00
7. Always has	26	3	9 54	3 24
8. Directions	14	3½	5 00	1 72

The subtests in order of application, time allowances, number of test-units, and crude scores and sigmas, are given in Table I. The whole group test took forty minutes for complete application.

INTERCORRELATIONS AND TETRAD-DIFFERENCES

Crude Scores.—Intercorrelations for subtests 1 to 8, with age, are given in Table II. The difference formula for *r* was used, for crude scores. The correlations are a result of checkings, both in the normal course, and at the instigation of the tetrad-differences themselves. We examine the correlations in terms of the Spearman Theory of Two Additive Factors.[2] The influence of age is neglected for the present: Compared with other disturbances to be considered, age effect is but slight. Table II provides the following tetrad data.

TABLE II.—PRODUCT-MOMENT CORRELATIONS FOR 1037 GIRLS, CRUDE SCORES, VERBAL SUBTESTS

Age	1	2	3	4	5	6	7	8
	1065	0947	0795	−0018	0822	−0018	−0686	0321
1		6408	5706	4289	5151	4408	5680	6110
2			6121	5691	5980	5770	5897	5799
3				4701	5871	5524	5724	5236
4					4556	5049	5010	514
5						5174	5654	5240
6							5600	5014
7								5859
8								

Mean of 210 tetrad-differences	0.0323
Conventional observed pe (Mean × 0.8453)	0.0273
Theoretical PE (16A^2)	0.0104

We thus find error in the tetrads, over and above that attributable to sampling.

The tetrads involving subtest No. 1 have observed pe of amount 0.0353, and supply some of the largest of the 210 tetrad-differences. This subtest was introduced as a "shock-absorber," and was in view of the girls during the preliminary explanations of testing routine—it is probably potent as a disturber of tetrads because of a bias for "speed." It seems that we cannot take liberties of the kind granted to subtest No. 1, without having disturbances resulting in tetrads which involve the subtest. In the circumstances it is perhaps legitimate to omit subtest No. 1 from all further deliberations and this, in any case, is the procedure adopted: Its retention would not greatly effect the data obtained from consideration of the remaining seven subtests. The latter, subtests 2 to 8, provide 105 tetrad-differences:

Mean of 105 tetrad-differences	0.0229
Observed pe (Mean × 0.8453)	0.0194
Observed sigma ($0.6745\sqrt{\Sigma t^2/n}$)*	0.0187
Theoretical PE approximately	0.0104

* Where t stands for tetrad-difference, n their number

We note error of amount about 0.015 in excess of that that can be attributed to sampling.

Standard "Normal" Scores.—Following the procedure already used for the non-verbal subtests, the crude scores used for the correlations of Table II were converted to the standard "normal" scores given previously.[7] Subtest No. 1 was not rescaled. After rescaling, all the subtests have the same "normal" distribution of scores. The new correlations for subtests 2 to 8 so rescaled are given in Table III, age being neglected. The mean intercorrelation is now 0.5124, compared with 0.5456 for the corresponding r's of Table II. Table III provides the following data:

TABLE III.—PRODUCT-MOMENT CORRELATIONS FOR 1037 GIRLS STANDARD "NORMAL" SCORES; VERBAL SUBTESTS

	2	3	4	5	6	7	8
2		6013	5589	5085	5076	5845	5883
3			4623	5024	5266	5500	5106
4				4474	4593	5019	5046
5					5550	5745	5121
6						5697	5220
7							5929
8							

Mean of 105 tetrad-differences	0.0249
Observed pe (Mean × 0.8453)	0.0211
Observed sigma ($0.6745\sqrt{\Sigma t^2/n}$)*	0.0202
Theoretical PE approximately	0.0104

* Where t stands for tetrad-difference, n their number

It is apparent, then, that the rescaling has not freed the table of the excess error observed for Table II. Our immediate object is to search for acceptable disturbances in these tetrads Throughout we shall make use of Table III, the correlations being there free from scaling anomalies

CONSIDERATION OF THE RESIDUAL ERROR

Age is of but slight effect on the tetrads and, from our experience with the non-verbal subtests, we take the correlational calculations to be too accurate to allow of calculational mistakes being suggested as an explanation of the error obtained for the tetrads of Table III What has been said in a previous paper[7] concerning calculation mistakes holds equally well here: There must be *some* error attributable to calculational mistakes, but we take it to be slight; and the various correlations should be relied upon even if it is found that the tetrad-differences are slightly inaccurate, because the correlations receive most attention. The calculations for Table II were quite different, from beginning to end, from those for Table III: The crude scalings were satisfactory approximate "normal" distributions, so that the correlations for Table II should differ but little from those for Table III As we see, the correlations are in fact but little different. Such considerations, and similar details to be observed in the course of our work, together with the help provided by instituting recheckings of calculations at the investigation of the tetrads themselves (a correlation associated with large tetrad-differences is first suspected of calculational mistakes, and rechecking is made for that correlation), tend to verify our acceptance of the various correlations as being reasonably free from calculational mistakes

We can isolate no single specificality in the case of Table III (or II), *i.e.*, none showing uniquely, as in the case of the $r_{\text{II},\text{VIII}}$ correlation or the non-verbal subtests.[7] We therefore proceed to examine Table III in the light of influences already known to us, so that, by eliminating some, a limited number may be left for final consideration

Similarity of Relations —The Classification, Opposites, and Analogy subtests (3, 5, and 6 respectively) involve very similar "likeness," or "unlikeness," eduction—the classification test-units can be answered in terms of "not-similar," or "opposite," the analogy test-units involve both "opposites" and "similars," while the opposites test-units likewise sometimes may imply "similars " Thus, for example, the analogy test-unit is constructed of "opposites" in the following sample:

black : white :: dark: ___?___

Now these three subtests have been found by Davey[1] to entail specificality which, not unreasonably, was attributed to too great similarity of the relations involved in the test-units. It is probably, then, that the same effect enters into our subtests. We can free our data of such an effect by omitting all tetrads involving r_{35}, r_{36}, and r_{56}, leaving for consideration the tetrads that cannot be disturbed by the probable specificality. The process leaves 57 tetrads, with observed pe (conventional, *i.e.*, observed mean × 0.8453) of value 0.0181, the theoretical PE being approximately 0.0104. Hence, even were specificality acceptable for these correlations, the residual error for Table III is just slightly reduced by eliminating the disturbance.

"*Idiosyncrasies.*"—Of other possible disturbers of tetrads previous mention has been made of test-constructor's idiosyncrasies.[4] Thus, a battery of Thorndike CAVD subtests compared (by means of the tetrad technique) with, say, Spearman oral subtests, shows specificality in the one set of subtests relative to the other.[6] The specificality may be due to the fact that the Spearman test-units are orally presented; or to scholastic influences in the Thorndike test-units (showing in Arithmetic, Word-ability, Understanding of paragraphs, etc.); or, indeed, to both or other influences.

Now of the verbal subtests in Table III, Nos. 2 and 5 were largely Thorndike CAVD test-units (sentence-completion and opposites), while the others were of my own construction (although following well-known patterns). The tetrads for the five subtests of my construction (3, 4, 6, 7, and 8) have value as follows:

Mean of 15 tetrad-differences . .	0.0148
Mean × 0.8453 (observed pe) .	0.0125
Theoretical PE approximately	0.0105

It is of interest to note that the largest tetrads among these 15 are associated with the correlation r_{36} (considered above): Omission of r_{36} gives 9 tetrads with value.

Mean of 9 tetrad-differences .	0.0131
Mean × 0.8453 (observed pe)	0.0111
Theoretical PE approximately . .	0.0105

(If the comparable correlations for crude scores are considered the tetrads still show error of amount 0.015 when r_{36} is omitted: So that if we accept the calculations, error of amount 0.015 appears to arise from scaling anomalies, a result in accordance with conclusions drawn in the previous paper.[7]

The subtests 2 and 5 may be disturbed by difficult or unfamiliar words; subtests 3, 6, 7, and 8 contain only what must be taken to be familiar words with "concrete" meaning (see the various samples already given) and verbal subtests less influenced by critical word-ability could scarcely be constructed Subtest 4 involves somewhat complex sentences, with abstract structure.

The omission of tetrads involving r_{25} would free the tetrads for Table III of any influence characterizing these two subtests *alone* But residual error of amount 0 019 remains if r_{25} alone is omitted from the tetrads for Table III, and if, in addition, we omit tetrads involving r_{35}, r_{36}, and r_{56}, so making allowance for possible "similarity of relations," this residual error becomes 0.018 (over and above sampling error) Thus idiosyncrasy for subtests 2 and 5 provides no obvious specificality relative to the other subtests.

The author's work with various group tests—such as the Otis, Spearman, National, Thorndike, etc —has shown the potency of influences entailed in different test-units constructed by one and the same psychologist:[6] Above, we suggest, is an example of the same phenomenon. Had the author constructed the eight verbal subtests himself, (in passing we note that subtest No. 1 was taken from work by Slocombe[5]) he is of opinion that the eight might have shown agreement with the tetrad criterion as satisfactory as that shown by 3, 4, 6, 7, and 8. Subtests 2 and 5, acting as "reference values," show that disturbances are latent in the five subtests

"*Speed Preference* "—An example of speed preference is given in the previous paper,[7] for subtests II and VIII We found that the effect could be controlled in terms of the errors made in subtest VIII

Subtests 2, 4, and 8 (and 1) were free from errors Each test-unit was either fully correct or not attempted Subtests 3, 5, 6, and 7, show frequent errors Sheer guessing, the scores indicated, was rare. We judge introspectively, and from the prevalence of mistakes made by the testees, that subtests 3, 6, and 7, are most likely to entail a "speed" effect Subtest 5 is ostensibly similar to 3, 6, and 7, and may be judged to entail "speed preference", but the test-units in subtest 5 have qualities that do not lend themselves to "speed preference " Furthermore, r_{35} and r_{56} have received prior notice in terms of similarity of relations, and the inclusion of these correlations in the tetrads about to be considered is open to criticism on that account The influence of "speed" will be most noticeable in tetrads of the type

$$r_{s_1s_2} \cdot r_{q_1q_2} - r_{s_1q_1} \cdot r_{s_2q_2} = F \qquad (1)$$

(s stands for a subtest involving "speed preference," such as 3, 6, and 7; and q for a subtest taken to be satisfactory for the speed-quality functioning.)

There are 36 such tetrads for the s subtests 3, 6, 7, and 5, taken together with the q subtests 2, 4, and 8. These have observed pe 0.028—quite a significant result. Omitting these 36 tetrads from the full 105 for Table III leaves 69 tetrads, with value

Mean of 69 tetrad-differences. .	0.0206
Mean × 0.8453 (conventional pe) .	0.0174
Theoretical PE approximately. . . .	0.0105

But if subtest 5 is not taken to be "speed" biased or, if prior regard is taken of the similarity of relations noted for r_{35} and r_{56} (and probably r_{36}), then the mean of 18 tetrads of form (1) for s subtests 3, 6, 7, and q subtests 2, 4, 8, is 0.024: And elimination of these leaves the tetrads for Table III undiminished. We see, then, very little clear evidence that "speed preference" has entered significantly in our subtests, judged, that is, among themselves. The evidence, of course, might be otherwise if different "reference values" are employed.

"Propinquity" Influences.—These influences may be attributed to objective or subjective fatigue (especially when the testing time is lengthy, taking two hours or so) as "habituation," and as "end-spurt." Calculation mistakes may be suspected too, especially in r_{12}—which is usually the first correlation calculated. The influences show in tetrads of the type:

$$r_{12} \cdot r_{78} - r_{17} \cdot r_{28} = f$$

Control of such influences was attempted in our work. Subtest 1 has been omitted from calculations partly because it served as an "habituation," or "shock," absorber, the testing time was reasonably short; subtest 7 was quite unlike subtest 8 in routine requirements, so tending to break any "set" toward extra effort in the last two subtests; every endeavour was made to dispel initial indifference, apprehension, or "speed preference."

The observed conventional pe of all tetrads involving r_{78} is 0.0244. But the largest of these tetrads have r_{35} and r_{56} likewise on the left-hand side of the above type of equations, and omission of tetrads involving these correlations leaves 16 tetrads, now with pe 0.0205. Omission of all tetrads involving r_{78}, r_{35}, and r_{56}, leaves 51 for Table III, with pe 0.0168. However, what little evidence may be got from these results for r_{78} is offset by the fact that no disturbance is apparent for r_{78} when the author's subtests alone are considered (see section on *Idiosyncrasies*). The data thus give no indication of disturbances attributable to propinquity effects.

The Influence of Testing by Groups.—This has been suggested by Professor Spearman[2] as a likely source of error in tetrads. In the case of the present verbal material a check of the influence can be made by calculating separate correlations for each group of girls tested, averaging the correlations so obtained. For our 1037 girls we had 21 testing groups. The calculation mistakes entailed in working out such a large number of correlations might outweigh the error to be expected as attributable to group testing. It is possible that, by calculating separate correlations for each testing group, we might arrive at intercorrelations giving sampling error values to the tetrads for Table III. But we would be concerned with the slight specificalities under consideration for Table III, the whole being relative to the verbal subtests themselves.

We note that no error need be (nor can be) attributed to group testing anomalies in the case of the non-verbal subtests. And the largest error available in the case of the verbal subtests can be of the order 0.015 only. In the case of the five subtests constructed by the author, 3, 4, 6, 7, and 8, there need be no error attributable to group testing anomalies.

But there can be no doubt that group testing, on occasion, may introduce large disturbances in tetrads. Thus, correlations may be much influenced if one class, having been taught geometry, is given a test involving geometry, while other classes, given the same test, have received no such instruction in geometry. It may be taken, we suggest, that our non-verbal subtests would be most open to influences of this kind (consider, for instance, the subtests III, VI, and VIII as given in the previous article[7]). Except in terms of vocabulary instruction, such influences cannot readily be taken to enter the verbal subtests. Previous practice with similar subtests would not cause disturbances in the tetrads, if *all* the subtests received the practice, or, in effect, received the practice. Group testing may introduce disturbances in other ways; for example by mistakes in timing subtests, or through faults in testing technique for particular subtests for some testing groups. All the testing was done by the author, and suggestions of timing or test application or marking faults can be discountenanced.

A procedure to be described in the next section might be expected to negate broad influences that may arise because of group testing. The procedure leads to no material diminution of the error in the tetrads for Table III.

The Influence of School and Class.—As has been reported in the previous paper,[7] the girls for our work were drawn from eleven schools, and from Standards IVA to VIIB in these schools (see Table II[7]). Certain schools were by repute of better standing than others. We could determine intercorrelations for the subtests and reputed school standing. Similarly, a point scale could be made for school standard. The influence of school and school standard could then be partialled out. Instead of making two sets of correlations necessary, our purpose can be served by combining both reputed standing and school standard (class) into one measure. A 16-point scale was devised (hereafter called the C-measure), the foundation being the order of class (IVA to VIIB) within which girls of classes with high reputation were accommodated. The various Standards IV were scaled 0, 1, 2 etc., where the reputed "poorest" school was scaled 0. The score was the same for every girl in a particular class: And it depended upon opinion given by teachers, upon scholarship returns, and, objectively, upon the school class. We believe the C-measure to be a satisfactory measure of the relative schooling and environmental influences exerted on the girls.

To Table III we can now add the following correlations of the subtests 2 to 8 with the C-measure: 0.4598, 0.4291, 0.3130, 0.3659, 0.4163, 0.4451, 0.4110, respectively.

The partial correlations, for the C-measure partialled out were next calculated, and the correlations provided tetrads with the following value:

Mean of 105 tetrad-differences, Table III, with C-measure partialled out.	0.0253
Observed pe (Mean × 0.8453)	0.0214
Theoretical PE approximately	0.0105

Furthermore, we can try out all the various possible influences discussed in the previous sections, using the partial correlations instead of those given in Table III. In all cases the conclusions arrived at are in no way different from those arrived at for Table III The disposition of error in the tetrads still shows r_{35} and r_{56} as sources of large differences; a try out of the theory of "speed preferences" gives much the same results; and the five subtests constructed by the author again have good agreement with the tetrad criterion. Thus, it is not in terms of a C-measure that residual error in the tetrads for the verbal subtests can be explained.

The Influence of "Verbality"—Ostensible characteristics of the verbal subtests are as follows

(*a*) Subtests 2, 4, and 8, involve phrases and sentences or short paragraphs.

(*b*) Subtest 5 (and 1 to a lesser degree) critically entails difficult words—for instance, test-units of this type.

"airy uniform blithe diversified"

(*c*) Subtests 3, 6, and 7, make use of simple words, all well within easy understanding of the girls tested (*i e*, according to their g-ability—we do not suggest that the dullest girl tested had knowledge of all the words in these subtests) (It is this easiness of vocabulary, and simplicity of the relations involved, that makes subtests 3, 6, and 7 particularly open to "speed" influences)

(*d*) Subtest 8 also involves easy words, equally understandable by the girls, but we cannot readily take it to be "speed" biased

(*e*) Subtests 2 and 4 are not so critical for vocabulary as is subtest 5 Subtests 2 and 4, however, are superficially much alike—both are essentially "sentence completion"

(*f*) A generalization from influences of the kind just passed in review ((a) to (e)) is got in terms of the fact that for any given fundaments—of words, or phrases, or other linguistic structure—a critical matter may be (1) either that the word or phrase, etc has to recall meaning, or (2) that meaning (already educed) has to recall words or phrases, etc Thus, *reproduction*, recall of a word or phrase or meaning, may be critical in some test-units But, education, not reproduction, characterizes the universal g-factor, and critical reproduction may thus possibly act as a disturber of tetrads

What, then, has our data to offer concerning the above characteristics and influences? Obviously it will be of greatest value to refer the verbal material to the non-verbal subtests, keeping in mind the above possible influences This, however, is to be our concern later For the present we note the following facts

Tetrads of type (1) for s as subtests 3, 6, and 7 (taken two at a time) and q as subtests 2, 4, 5, and 8 (two at a time), have observed conventional pe 0.0201, while the observed pe for all the other tetrads of the table is 0 0216 If we allow for disturbances attributed to "similarity of relations" the observed pe for tetrads of type (1) becomes 0 0174, compared with 0.0178 for the rest of the tetrads of the table There is no evidence, then, for specificality in subtests 2, 4, 5, and 8, relative to 3, 6, and 7, *i.e.*, nothing is observed of the nature of a "speed preference" for subtests 3, 6, and 7, and nothing in terms of "verbal ability" in subtests 2, 4, 5, and 8 (for difficult vocabulary and understanding of connected discourse). This conclusion, of course, is relative to the verbal subtests amongst themselves. The results do not seem to promise assistance by offering explanations of the excess error obtained in the tetrads for Table III.

In terms of the verbal subtests alone, it would be making too much of our material to enter into the possible influences of *reproduction* on the tetrads. Should a sufficiently broad specificality be observed in the verbal subtests relative to the non-verbal subtests, which receives no acceptable explanation in terms of influences such as those of "speed preference," "idiosyncrasies," "group testing," and the like, then a most important source of disturbances of tetrads might well be *reproduction* effects All this, however, is a matter for future consideration.

RÉSUMÉ AND CONCLUSIONS

We have applied the Theory of Two Factors to the intercorrelations for Table III, and find that the observed error in the tetrads is slightly in excess of that expected from sampling error alone

Previous work confined to verbal subtests has shown some agreement with the tetrad-difference criterion, but most of the previous data are for small populations only, where the sampling error is large in comparison with other errors, so swamping them. When, however, the sampling error is small (as in our work) some observed excess error, of the kind that we have found, is to be anticipated and, indeed, is demanded by the Theory of Two Factors The observed excess error for Table III, then, is not incompatible with the Theory But the elimination of such small excess errors would only slightly reduce the g-saturation of the correlations.

Our problem has been that of seeking acceptable explanations for the excess observed error in the tetrads for Table III We have applied to our data certain theories of disturbances of tetrads that have been found of explanatory value in previous work, trying, in particular, the influences of age effects, calculation mistakes, "speed preferences," "propinquity," tester's idiosyncrasies, school and class, group testing anomalies, and "similarity of relations." Of the various suggested disturbers of tetrads, that of "similarity of relations" for the Analogy, Opposites, and Classification subtests appears to be the most discernible, but it is not at the root of the major part of the observed excess error. "Speed preference," age, school and class, and superficial verbal likenesses of the subtests, give no clear indication of being appreciably contributory to the excess error. We are in some doubt about the magnitude of error attributable to group testing anomalies. Error is perhaps likely, but the non-verbal subtests[7] gave no indication of being so influenced, and it is perhaps among these subtests that we might most expect such group testing effects to be disturbers of tetrads. We find, then, that none of these suggested probable disturbers singly can be taken to account fully for the observed excess error in the tetrads for Table III. The sources of error so far brought forward appear to be inadequate to explain the whole of the excess error

We might accept, provisionally, a theory that a sum of many small disturbances of the kind passed in view in the course of our paper may be taken to explain the observed excess error. This would mean that elimination of all the suggested small disturbances would but slightly reduce the g-saturation of the correlations. For further light on this matter, however, we must look to comparison of the verbal and non-verbal subtests, a study that we are to report in the next paper.

REFERENCES

1. Davey, C M A Comparison of Group Verbal and Pictorial Tests of Intelligence *British Journal of Psychology*, Vol XVII, 1926, p. 27
2. Spearman, C · "Abilities of Man Their Nature and Measurement" The Macmillan Co , 1927
3. Spearman, C Factor School of Psychology, Part XX *Psychologies of* 1930
4. Spearman, C Disturbers of Tetrads Scales *Journal of Educational Psychology*, Vol XXI, 1930, p 559
5. Slocombe, C Thesis, Library of University of London
6. Stephenson, W Thesis, Library of University of London
7. Stephenson, W . Tetrad-differences for Non-verbal Subtests *Journal of Educational Psychology*, March, 1931

FOUR TYPES OF EXAMINATIONS COMPARED AND EVALUATED

ALVIN C EURICH

University of Minnesota

Since objective examinations are used so extensively to measure achievement in college courses, it may no longer be necessary to refer to them as new-type examinations. This does not imply that such measures of accomplishment are fully evaluated nor does it even suggest that diligent scrutiny of their results has become anachronistic. In fact the greater use of these examinations makes imperative even more extended studies of their relative validity and reliability as well as the inter-relationship that exists between them. Such investigations should be conducted in a rather wide variety of specialized courses now taught in colleges and universities. The well-known work of Wood[1] at Columbia University has done much to stimulate a widespread interest in evaluating various types of achievement examinations that are now in vogue. The particular investigation reported within these pages is an outgrowth of that interest. Its purpose is to evaluate the essay, completion, multiple-choice, and true-false examinations when each type covers exactly the same subject-matter.

METHOD

As far as can be determined the unique feature of this investigation rests in the method employed to evaluate the particular types of examinations used. In constructing the tests a traditional essay-type examination was first prepared to cover the salient points of the subject-matter Following the formulation of the questions the answers were written out in detail and the total number of items to be included in each question was thereby determined The next step in the procedure was to construct a completion examination covering exactly the same material as that embodied in the essay type. To be assured of this, the completion statements were made up from the answers to the essay questions. A multiple-choice examination was next constructed in the same manner; and following this, a true-false examination. Thus the completion, multiple-choice, and true-false tests were each derived from the correct responses to the essay-type questions.

---

[1] Wood, Ben D. "Measurement in Higher Education" World Book Co, 1923

This procedure was followed in two experiments at the University of Minnesota.

The first of these was conducted with a group of students enrolled in a course in statistical methods during the first term of the summer session of 1927. In the class, there were one hundred seven seniors and graduate students of the College of Education, ninety-nine of whom were present on the day the tests were administered.

In the second experiment, the group consisted of one hundred six juniors and seniors enrolled in educational psychology during the first term of the summer session of 1929.

In both cases the only knowledge which the members of the class had concerning the experiment was that they were to take a two-hour midterm examination. They were told that this examination was to consist of four parts. Part I, an essay examination, Part II, a completion examination, Part III, a multiple-choice examination; and Part IV, a true-false examination. They were further informed that the examination would cover all the material in the course up to the middle of the term.

The total number of points in each type of test and the amount of testing time allotted are included in Table I for the first experiment and Table II for the second. The examinations used in Experiment II were somewhat longer than those in Experiment I although the amount of time allotted to each was greater only in the case of the true-false test. The experience with the first investigation served as a basis for this adjustment. The tests are listed in the tables in the same order as they were given.

TABLE I.—TESTING TIME, TOTAL POINTS, MEANS, AND VARIABILITY OF SCORES ON FOUR TYPES OF TESTS USED IN EXPERIMENT I. $N = 99$

Test	Testing time in minutes	Total points	Range of scores	Mean	PE	SD	PE	CV
Essay	35	69	20– 61	39 27	± 62	9 14	± 44	23 27
Completion	30	94	25– 78	49 16	± 78	11 44	± 55	23 27
Multiple-choice	20	39	11– 38	29 41	± 33	4 91	± 24	16 70
True-false	15	50	28– 46	39 35	± 23	3 34	± 16	8 49
Composite of four tests	100	252	109–217	154 89	±1 57	23 20	±1 11	14 98

The scoring of the essay test was somewhat more objective than is usually the case Each student's paper was compared with the complete set of answers previously worked out, and only those points which appeared on the master copy were given credit For the completion and multiple-choice examinations, the number of correct responses was considered as the score On the true-false test the score was determined by subtracting the number of wrong responses from the number right

TABLE II —TESTING TIME, TOTAL POINTS, MEANS AND VARIABILITY OF SCORES ON FOUR TYPES OF TESTS USED IN EXPERIMENT II $N = 106$

Test	Testing time in minutes	Total points	Range of scores	Mean	PE	SD	PE	CV
Essay	35	118	12– 70	40 80	± 80	12 15	± 56	29 78
Completion	30	96	16– 79	55 28	± 84	12 90	± 60	23 34
Multiple-choice	20	69	20– 62	48 44	± 48	7 35	± 34	15 17
True-false	20	75	5– 61	39 25	± 75	11 40	± 53	29 04
Composite of four tests	105	358	62–258	171 33	±2 37	36 20	±1 68	21 13

RESULTS

In the first experiment the four tests varied in their differentiating capacity as shown in Table I. An examination of the figures in the appropriate column reveals the fact that the standard deviations on the essay and completion tests are larger than on the multiple-choice and true-false examinations. The same is true of the coefficients of variability This difference is without question due to the nature of the particular examinations used rather than due to the type of test. The variability measures for Experiment II (Table II) show that the essay, completion, and true-false examinations appear to differentiate the students to approximately the same extent. Had the multiple choice test been made longer it is very probable that it would have differentiated the students equally well.

In order to determine the extent to which the four types of tests measured the same degree of achievement, their intercorrelations were found The coefficients for both experiments are given in Table III Considering the tests as given in the first experiment it is evident that the highest intercorrelation appears between the multiple-

TABLE III.—INTERCORRELATIONS OF TESTS

	Experiment I			Experiment II		
	r	PE	Corrected for attenuation	r	PE	Corrected for attenuation
Essay and completion	44	± 05	02	80	± 02	1 20
Essay and multiple-choice	47	± 05	67	63	± 04	97
Essay and true-false	30	± 06	66	55	± 05	89
Completion and multiple-choice	57	± 05	80	71	± 03	92
Completion and true-false	37	± 06	68	63	± 04	85
Multiple-choice and true-false	44	± 05	81	65	± 04	90

choice and true-false types ($r = .57$) The lowest exists between the essay and true-false examinations ($r = .30$) When corrected for attenuation, the multiple-choice test correlates about equally as high with the completion (.80) as with the true-false type (.81) In the second experiment all the intercorrelations are higher The coefficient of greatest magnitude is found for the essay and completion tests (80) The lowest correlation obtains between the essay and true-false tests (.55) as in the first experiment. When corrected for attenuation the coefficients in Experiment II range from 85 to 1.20 [1] These high coefficients indicate that if a reliable test is constructed one type is probably as adequate for measuring the acquisition of information in a course as any of the other three types The fact that the intercorrelations are lower in the first experiment than in the second is probably due to the lower validity and reliability of the first set of tests in comparison with the second, although the subject-matter of the tests may also be a factor.

To evaluate these facts further, correlation coefficients were calculated for the scores on each test and composite scores on the other three For example, the essay test was correlated with the composite score of the multiple-choice, true-false and completion examinations. The same procedure was followed for each of the other tests These correlations, which may be called validity coefficients, have been placed in Table IV. The values for the tests as used in the first experiment appear in the second column as follows: For the essay, .64;

---

[1] A corrected correlation that is greater than 1 00 must be interpreted to mean that the relationship is near unity

for the completion, 68; for the multiple-choice, 66, and for the true-false, 40. It was further sought to determine the validity of the tests in case each of the four types was made of equal length; that is, in case the testing time for each was sixty minutes Thus, the validity coefficients are estimated for the essay test when less than doubled in length, for the completion test when doubled, for the multiple-choice test when tripled, and for the true-false test when quadrupled To do this the formula for determining the validity of a lengthened test as given by Holzinger[1] was used. The coefficients obtained in this manner are as follows: Essay, 69, completion 73, multiple-choice, .74; and true-false, .54.

TABLE IV—THE RELATION OF EACH TYPE OF TEST TO THE COMPOSITE OF THE OTHER THREE TYPES

Tests	Experiment I			Experiment II		
	As constructed		If made sixty minutes in length	As constructed		If made sixty minutes in length
	r	PE	r	r	PE	r
Essay	64	± 04	69	76	± 03	84
Completion	68	± 04	73	82	± 02	.86
Multiple-choice	66	± 04	74	77	± 03	84
True-false	40	± 06	54	75	± 03	81
Estimated validity of composite of four tests	70			80		

The validity coefficients for the tests used in Experiment II were likewise calculated These are also inserted in Table V For the essay test as constructed, the validity coefficient is 76, for the completion, .82, for the multiple-choice, 77, and for the true-false, 75. When estimated for the tests if made of equal length, the coefficients are all of approximately the same magnitude They indicate, therefore, that the four types of tests used in these experiments have approximately the same validity (providing the criterion is an adequate one) The surprising implication of this fact is that the essay type of examination, when scored objectively, appears to be as valid as the other so-called objective examinations

---

[1] Holzinger, Karl J. "Statistical Methods for Students in Education" P 170.

The reliability coefficients presented in Table V were found by securing the Pearsonian r between the odd and even items of the tests and estimating with the Spearman-Brown formula. In both of the experiments the completion and multiple-choice examinations are most reliable In the first, the true-false test has the lowest reliability whereas the essay examination is least reliable in the second. When the reliability coefficients are estimated for tests of equal length, the completion and multiple-choice types retain their positions as the most reliable tests in both experiments The reliability of the essay examination in the first experiment does not differ markedly from these whereas in the second experiment, the essay examination appears to be considerably less reliable The reliability coefficient for the composite of the four tests of each experiment was determined by an adaptation of Spearman's formula for the correlation between sums or averages.[1] The coefficients show that the two batteries of tests have approximately equal reliability

TABLE V —RELIABILITY OF TESTS

Tests	Experiment I				Experiment II			
	As constructed		If made sixty minutes in length		As constructed		If made sixty minutes in length	
	r	PE	r	PE	r	PE	r	PE
Essay	00	± 04	70	± 03	56	± 00	69	± 04
Completion	72	± 04	84	± 02	80	± 03	89	± 02
Multiple choice	71	± 04	88	± 02	75	± 03	.00	± 01
True-false	41	± 08	74	± 05	69	± 04	87	± 02
Composite of four tests	86				89			

Another matter of concern was the relationship of these various types of examinations to intelligence The only intelligence test scores available for the group of students enrolled in statistical methods were those on the Miller Mental Ability Test, Form A. The coefficients of correlation between this test and the various types of achievement examinations appear in Table VI. The zero order coefficients show that the highest degree of relationship exists between Miller A test and the completion examination (r = 53), the lowest between the Miller A and the essay examination ($r = .33$). When corrected for attenuation, the coefficients of correlation between the Miller Mental Ability Test with the completion (.68) and with the true-false examination (.67) are approximately equal, whereas the lowest relationship is found with the essay examination.

---

[1] Kelley, T. L : "Statistical Method" P 198, Douglas, H R and Cozens, F. W : On Formula for Estimating the Reliability of Test Batteries *Journal of Educational Psychology*, Vol. XX, 1929, pp 369–377

[2] Miller, W S "Mental Ability Test" World Book Company

TABLE VI—RELATION OF TESTS TO INTELLIGENCE

Tests	Experiment I Miller Mental Ability Test			Experiment II Miller Hard Analogies Test		
	r	PE	Corrected for attenuation[1]	r	PE	Corrected for attenuation[1]
Essay	33	± 00	43	34	± 00	48
Completion	53	± 05	68	44	± 05	52
Multiple-choice	42	± 06	54	53	± 05	65
True-false	40	± 06	67	55	± 05	70
Composite of four tests	49	± 05	57	53	± 05	59

[1] The self correlation for the Miller A test is 86, and for the Analogies test, .90

In the educational psychology class the Miller Hard Analogies Test[1] was used as a measure of intelligence. With this test, which is better adapted to the college group than the Miller Mental Ability Test, the relationship between the essay examination and intelligence is also the lowest of those obtained It would seem, therefore, that students with relatively good intelligence are less apt to secure high scores in the essay examination than on the other three This may be due to the fact that the form of the essay type is much less like the form of intelligence tests than are the completion, multiple-choice, and true-false types.

After the students completed all four parts of the examination, they were asked to write out their order of preference for these types. These data are summarized in Table VII for Experiment I and in Table VIII for Experiment II. In both classes the multiple-choice examination is placed first by a larger percentage of students than any of the other types. The true-false is placed second. In the first experiment the plurality gives the essay examination the third position and the completion examination the fourth In Experiment II, however, the plurality gives the completion the third position and the essay, the fourth. It is clear that the multiple-choice and true-false examinations are preferred by the students to the essay and completion types.

---

[1] Not commercially available Prepared by W. S Miller

TABLE VII —STUDENT PREFERENCES FOR DIFFERENT TYPES OF TESTS

Experiment I

Choice	Type of examination							
	Essay		Completion		Multiple-choice		True-false	
	N	Per cent	*N*	Per cent	*N*	Per cent	*N*	Per cent
First	14	14.9	6	6.5	40	43.0	39	40.6
Second	14	14.9	5	5.4	34	36.6	39	40.6
Third	44	46.8	22	23.9	15	16.1	11	11.5
Fourth	22	23.4	59	64.1	4	4.3	7	7.3
Totals	94	100.0	92	99.9	93	100.0	96	100.0

TABLE VIII —STUDENT PREFERENCES FOR DIFFERENT TYPES OF TESTS

Experiment II

Choice	Type of examination							
	Essay		Completion		Multiple-choice		True-false	
	N	Per cent	*N*	Per cent	*N*	Per cent	*N*	Per cent
First	11	11.5	13	13.4	58	59.8	18	18.4
Second	9	9.4	16	16.5	27	27.8	44	44.9
Third	27	28.1	38	39.2	9	9.3	22	22.5
Fourth	49	51.0	30	30.9	3	3.1	14	14.3
Totals	96	100.0	97	100.0	97	100.0	98	100.1

Further analysis was made of these data by tabulating the order of preference for the fifteen individuals securing the highest score and for the fifteen securing the lowest score on the intelligence tests Table IX presents the figures for these selected groups in the statistical methods class while Table X presents comparable data from the second experiment Although no generalizations can be derived from these two tables, it appears that both those of low and high intelligence prefer the multiple-choice and true-false tests more than they do the other types. In the first experiment there is a tendency for those lowest in intelligence scores to prefer the essay examination more than do those who received the highest ratings. This tendency is not as clear in the results from the second experiment In both classes, there is a slight tendency for those with the highest intelligence ratings

TABLE IX.—TEST PREFERENCES FOR THE FIFTEEN HIGHEST RANKING STUDENTS AND FOR THE FIFTEEN LOWEST RANKING STUDENTS IN INTELLIGENCE

Experiment I

Choice	Type of examination							
	Essay		Completion		Multiple-choice		True-false	
	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest
First.	0	5	1	0	7	4	7	6
Second	3	2	2	0	6	6	4	7
Third	7	7	4	3	2	4	2	1
Fourth	5	1	8	12	0	1	2	1
Totals	15	15	15	15	15	15	15	15

TABLE X.—TEST PREFERENCES FOR THE FIFTEEN HIGHEST RANKING STUDENTS AND FOR THE FIFTEEN LOWEST RANKING STUDENTS IN INTELLIGENCE

Experiment II

Choice	Type of examination							
	Essay		Completion		Multiple-choice		True-false	
	Highest	Lowest	Highest	Lowest	Highest	Lowest	Highest	Lowest
First	3	3	1	0	10	7	1	5
Second	2	2	3	1	2	5	8	7
Third	2	4	6	5	1	3	6	3
Fourth	8	6	5	9	2	0	0	0
Totals	15	15	15	15	15	15	15	15

to prefer the completion examination more than do those individuals who rank lowest

Before summarizing the results of this study, a possible explanation might be given for the lack of agreement between the results reported here and those reported by other investigators who have considered the relative merits of the essay examination. In all probability the reason for the relatively high rating of the essay examination in this study is the objective manner in which it was scored. The procedure followed in scoring these tests is definitely comparable to the method of scoring the so-called objective examinations Consequently, this evaluation of the essay examination does not serve as a refutation for the numerous arguments which appear in the literature opposing it. It merely intimates that an essay-type examination may be so given and scored as to compare favorably with the objective examinations. The labor involved in such a procedure, however, is so great as to make it almost prohibitive for large classes

SUMMARY

1. Four types of examinations, each covering exactly the same subject-matter, were given to a class in statistical methods in education and a class in educational psychology. The nature of the four types, as well as the order in which they were given, was as follows· Part I, essay; Part II, completion, Part III, multiple-choice; Part IV, true-false.

2 The intercorrelations of the tests in the course in educational psychology suggest that if reliable tests are constructed, one of the four types used is probably as adequate as any of the other three for measuring the amount of information which the members of the class have accumulated The evidence for this suggestion is not as clear-cut in the class studying statistical methods.

3 If the composite score on three types of examinations is used as the criterion for estimating the validity of the fourth type, the results indicate that the four types of tests have approximately equal validity.

4 In both experiments the completion and multiple-choice tests prove to be most reliable In the second experiment the reliability of the true-false examination is not much lower

5. The correlation between intelligence and the essay examination is lower than between intelligence and either of the other three types

6. Considering the choice of all the students in the two classes, the multiple-choice and true-false tests are preferred to the other two types studied.

7. There is a tendency for the highest ranking students in intelligence to prefer the completion examination more than do the lowest ranking students. In the first experiment this latter group prefer the essay test more than do the members of the highest ranking group

AN EXPERIMENT DESIGNED TO TEST THE VALIDITY OF A RATING TECHNIQUE

THEODORE NEWCOMB

Western Reserve University

I

The validity of ratings of behavior, provided that two conditions are met, has been fairly generally accepted These conditions are that there be several competent judges,[1] and that they have ample opportunity to observe the behavior being rated. For the past two summers the writer has been in a situation where both of these conditions were adequately met Data were gathered of such a nature that ratings could be compared with more objective measures of the same behaviors, and enough raters were available so that a study of the uniformity of the ratings could be made.

The subjects for each of the studies here reported were thirty problem boys who had been sent for study and treatment to a summer camp maintained through the cooperation of Western Reserve University and the Child Guidance Clinic of Cleveland, Ohio The boys remained for a period of five weeks where they were under the constant observation of a psychiatrist and of six or more counselors trained in psychology and mental hygiene.

The first summer, in connection with another experiment, the attempt was made to validate certain "objective" measures of problem boy behavior by comparing them with ratings on the same behavior The results were such as to make the experimenter conclude that the ratings were themselves being validated The evidence, however, was rather meager, and it was decided to gather fuller data concerning the same problem.

II

During the first summer a daily record was kept of "specifically remembered incidents" included under twenty-six different behaviors.[2]

---

[1] Note particularly Rugg, H O. Is the Rating of Human Character Possible? *Journal Educational Psychology,* Vol XII, p 425 and Vol XIII, p 93 Rugg concludes that there should be at least three judges who have had a fair chance to observe subjects in the behaviors being rated

[2] For a fuller description of this experiment, together with a description of statistical techniques used, see the writer's "Consistency of Certain Extrovert-introvert Behavior Patterns in Fifty-one Problem Boys" Teachers College Contributions to Education, No 382

Each boy's record each day was kept on a form similar to the following sample·

How was he about getting up in the morning?
  Got up before rising hour
  Promptly when called
  Dallied, but on time for breakfast.
  Too late to be on time for breakfast

Records were kept for each boy by his own counselor, who had spent all or most of the day with him. Besides these daily records, the experimenter recorded some 8500 incidents of the above twenty-six and other similar behaviors Then, at the end of the camp period, ratings were obtained from each of seven men on the frequency of these same twenty-six behaviors Inasmuch as the observers had played, slept, and eaten with the boys nearly constantly for five weeks, a fair degree of accuracy in the ratings might be expected.

Six of the twenty-six behaviors measured, however, were of such a kind that no subject had been observed by any but one man—his own tent counselor. Nevertheless, each of the seven observers was asked to give ratings on these six behaviors also, in terms *of supposed* or imagined frequency. Six of the seven ratings, in other words, were guesses

It was not surprising to find a mean correlation between daily record scores and ratings, of 400 ± .102 for the twenty-six behaviors The range of correlations was from .02 to .73. What was somewhat surprising was that the mean correlation for the six guessed behaviors (*i e.*, between daily records and ratings) was fully as high as for the twenty observed behaviors The mean for the former was .451 ± 098, and for the latter, .396 ± .104. The difference is not statistically significant

The accuracy of the ratings became even more questionable when, as measures of uniformity, standard deviations among the seven ratings for each behavior were computed The mean *SD* for the twenty observed behaviors was 88 (the ratings were on a scale of 1 to 5), and for the six guessed behaviors, .81. Again the difference is not significant, and it is evident that the raters agreed no more closely about frequently observed behaviors than about behaviors which they had never seen.

III

A year later there was a similar group of boys at the same camp Somewhat different means of objective recording of behavior were used, as will be described. More extensive ratings were also obtained, and the problem of studying their relation to frequency of observation was directly attacked this time, rather than stumbling on to it as a side-issue

As before, definite types of behavior were being studied, for example, seeking adult attention. Cards were mimeographed, on which were listed all the common ways in which boys were actually found to seek adult attention When an incident of this sort was observed, the card was checked with the appropriate behavior, at that time or as soon after as was possible. Seven or eight observers who were constantly on the watch for these specific behaviors could be presumed to get a pretty fair sample of all such incidents that occurred There follows a sample card, showing one of the seven kinds of behavior that were being studied

SEEKING ADULT ATTENTION

Date
Time
Observer

Name of boy
Hiking alone with counselor
Voluntarily working alone with counselor
Hanging around counselor alone
Volunteering in task for counselor
Hanging around counselors' shack
Complaining to counselor of ailment
Tattling
Special invitation to counselor
Antics to annoy or draw attention of counselor
Demonstration of affection for counselor
Playing up to visitors
Others (indicate on reverse side)

Altogether some fifty specific behaviors were being thus recorded day by day. Naturally, some of these were found to occur with great frequency, and others to be comparatively rare. This matter of frequency with which the behavior is being reported may therefore be studied in relation to the agreement between the two measures of the same behaviors

It was desired to study not only the effect of frequency of reporting upon the agreement of the two measures, but also that of mere frequency of occurrence of behaviors not being recorded These problems, added to that of the previous summer, *i e*, a comparison of observed with unobserved behaviors, made a total of six kinds of behaviors to be studied, as follows.

1. Behaviors being recorded in large numbers.
2. Behaviors being recorded in small numbers
3. Behaviors not being recorded, but occurring frequently (as determined by the study of the previous summer).
4 Behaviors not being recorded, but occurring rarely (as determined by the study of the previous summer)
5 Behaviors observed in each boy by only one observer (his own counselor)
6 Behaviors never observed at all—imaginary situations

Ratings were therefore obtained on thirty behaviors, five in each of the categories just given. In Table I these behaviors are tabulated, together with the results of the ratings In the list as submitted to the raters, the behaviors were of course not thus tabulated, but occurred in random order. The number preceding each behavior indicates what this random order was.

Directions for ratings were as follows:

The following ratings are to indicate the *frequency* with which each behavior occurred Please rate as follows:

0 Never occurred at all.
1. Rare, very exceptional.
2. Occasional, not frequent
3. Fairly frequent.
4 Very frequent; prominent characteristic.

Indicate the degree of your rating by placing a plus after the figure if there is a fair degree of certainty, a minus if there is considerable uncertainty, and otherwise no sign

In the case of behaviors which you have not observed, make the best possible estimate of how frequently it *would occur*

Rate all the boys for behavior No 1, then all the boys for behavior No 2, etc.

TABLE I —BEHAVIORS ON WHICH RATINGS WERE OBTAINED

Group I Recorded in Large Numbers

5 Lying or sitting around alone
7. Bullying, with physical contact.
13. Assuming position of prominence at camp fire
15 Antics to annoy or draw attention of counselors
27. Interrupting or shouting above the group.

Group II. Recorded in Small Numbers

1. Being crowded out, avoiding being first
3. Taking the initiative in fighting or boxing
18. Complaining to counselor of ailment.
20 Moping or sulking alone
24. Profanity or obscenity

Group III Not Recorded—Frequently Observed

2 Coming late to meals
11 Expressing enthusiasm
17 Moving slowly while others hurried.
21 Singing aloud
23 Laughing

Group IV Not Recorded—Rarely Observed

10 Boasting, or loudly announcing intentions for the future
12 Continued chatting, loquaciousness.
14 Remaining quiet while others were noisy
19 Accepting discipline quietly
22 Failing to wash or comb

Group V Observed by Only One Counselor

4 Making his own bed neatly
8 Taking more than his share of food at table
26 Readily concurring with tent-mates' choice of activity
29 Doing more than his share of after-meal work at camp
30 Making noise in the tent before rising hour in the morning

Group VI Never Observed—Imaginary Situations

6 Stopping play at a previously agreed upon hour and coming in the house
9 Playing quietly in order not to disturb sick sister, if asked to do so
16 Minding the baby all Saturday afternoon, without leaving
25 Giving up attractive plaything to younger brother if he asked for it
28 Tidying up his own room at home

Thirty boys were thus rated on thirty behaviors by six observers. As one measure of the uniformity of the ratings, the *SD* of the six ratings was calculated, for each boy for each behavior Correlations between ratings and objective scores were also calculated for the ten behaviors being recorded (Groups I and II) Rating scores for these calculations were obtained by adding the six ratings for each behavior by each boy Objective scores were simply the total numbers of recorded incidents

Reliabilities of the ratings were also calculated for each behavior, by correlating the sum of the scores of three raters with the sum of the other three raters This serves as another measure of uniformity of the ratings The correlation between these measures of reliability and the *SD*'s for each behavior, is — 652, indicating that the two methods are measuring at least partly the same thing

The results of all these calculations are given in Table II.

TABLE II

Behavior number	Mean of thirty *SD*'s	Correlation between halves	Correlation with objective score	Number of incidents
Group I—Observed and Recorded Frequently				
5	7102	632	189	123
7	7218	730	542	98
13	8004	276	662	69
15	6642	816	652	133
27	8000	783	616	100
Mean	7305	647	538	104 6
Group II—Observed and Recorded Rarely				
1	7065	735	019	21
3	6956	861	349	19
18	6158	783	413	25
20	7119	928	158	11
24	6315	905	681	34
Mean	6782	842	390	22 0
Group III—Not Recorded—Frequently Observed				
2	7405	507		
11	7466	703		
17	8213	154		
21	8040	500		
23	7580	609		
Mean	7741	491		
Group IV—Not Recorded—Rarely Observed				
10	7182	635		
12	7165	754		
14	7977	778		
19	7166	801		
22	6943	850		
Mean	7289	763		
Group V—Not Observed Except by One Man				
4	7151	750		
8	6097	548		
26	7553	856		
29	7152	901		
30	8306	479		
Mean	7372	706		
Group VI—Never Observed—Imaginary Incidents				
6	8107	486		
9	7006	474		
16	6620	730		
25	7290	701		
28	6899	729		
Mean	7204	643		

Considering first the correlations between the two measures, for which calculations were possible only in Groups I and II, the mean coefficients are respectively .538 ± .087, and .390 ± .105. The *SD* of the difference of these two coefficients, using the formula

$$SD_{diff} = SD^2 \text{ mean I} + SD^2 \text{ mean II},$$

is .1254, the difference being equal to 1.10 sigmas, which can not be said to be statistically significant. As far as agreeing with the more objective measures is concerned, then, there is no significant difference between behaviors frequently recorded and those rarely recorded.

The relative uniformity of the ratings for each group of behaviors is fairly obvious from Table II. The mean *SD* for the groups of behaviors ranges between .6782 for Group II and .7741 for Group III. The *SD* of the difference between these two values (using the formula already cited) is .0282, the difference being equal to 4.6 sigmas. While this is a great enough difference to indicate probable significance, there is no significant difference between any other two groups of behaviors. The difference in this one case would seem to indicate that behaviors being recorded in small numbers are more uniformly rated than those being frequently observed and not recorded; but not significantly more consistently rated than those which were pure guesses. Groups I and III, which represent behaviors frequently observed, have a mean *SD* of .7395 and .7741, respectively, whereas Groups V and VI, representing behaviors never observed, have mean *SD*'s of .7372 and .7204, respectively. Neither the fact of being observed frequently, nor having been watched for and reported for a five-week period lent any advantage in uniformity of rating over the ratings which were mere guesses.

There was a barely significant difference between behaviors No. 24 and No. 17, those with the lowest and highest mean *SD*'s, respectively. The sigma of the difference is .0427, the difference being equal to 4.42 sigmas.

Mean *SD*'s for each boy were also calculated, *i.e.*, the mean of the thirty behaviors on which each boy was rated. There was somewhat more uniformity here than among the thirty behaviors. The lowest and the highest of these mean *SD*'s were, respectively, .6615 and .8163. The *SD* of the difference is .051, the difference being equal to 3.03 sigmas.

That this scarcely represents a real difference is further indicated by a calculation of the correlation between the number of recorded incidents for each boy (ranging from 26 to 295) and the mean *SD* for each boy This coefficient, representing the relation between the number of recorded incidents and the variation in ratings was 161 ± 120 This is probably not surprising, since the mere fact of not having recorded many incidents for certain boys would give the rater a fair idea of the frequency of the behavior being rated. Hence boys with few recorded incidents would tend to be as accurately and uniformly rated as those with many.

Very similar results are obtained by comparing the average reliability of the several groups of behaviors The only significant difference, again, is between that of Group II and Group III, the difference being equal to 4 1 sigmas We find the average reliability of the five guessed behaviors to be almost precisely the same as that of the five behaviors most frequently seen and recorded.

The mean correlation between one half of the ratings and the other half is .683 This gives a reliability, by the use of the Spearman-Brown formula, of .81 In order to secure a reliability of .90 it would be necessary to have 2.1 times as many raters, or about twelve. It should be added that the boys presented a wide range of behaviors, so that a fairly high degree of reliability should be expected

Another aspect of ratings which interested the writer was the comparative uniformity of ratings of specific behaviors and of generalized traits Ratings were obtained on but two such traits, as follows:

1 Degree to which each boy was an acceptable member to those of his own tent group.

2 Degree to which each boy was an acceptable member to the majority of the thirty boys in camp

Ratings were made on a scale similar to that used in the specific behavior ratings, and the mean *SD* for each trait was similarly calculated. The mean *SD*'s were, respectively, 6705 and 6520, for the above traits The self-correlations were 885 and 655 respectively. While these results are not significantly different from these shown in Table II, they show comparatively high uniformity, and further measuring of other traits might have shown a tendency to rate traits more uniformly. It is to be regretted that more data were not obtained at this point.

In the above calculations the degree of certainty which was indicated along with the numerical values for the ratings, has not been considered. Would the results have been different if the numerical values had been weighted according to the degree of certainty?

Table III reveals surprising differences in certainty among the six groups of behaviors. There is a steadily decreasing certainty from Group I to Group VI. The calculations shown in Table III were made by finding the algebraic sum of the number of boys rated with certainty (marked plus) and the number of boys rated with uncertainty (marked minus) for each behavior.

Now, with such decided differences in the certainty of the ratings for the various groups of behaviors, one might expect to find significant differences in uniformity of rating if certainty is taken into account. Calculations were therefore made according to the following system of weighting: Ratings marked certain were weighted three times; those marked neither certain nor uncertain were weighted twice, and those uncertain, once.

TABLE III.—DEGREE OF CERTAINTY OF RATINGS FOR EACH BEHAVIOR

Group I		Group II		Group III	
Behavior number	Certainty	Behavior number	Certainty	Behavior number	Certainty
5	41	1	18	2	−2
7	11	3	27	11	12
13	46	18	10	17	20
15	54	20	24	21	4
27	37	24	46	23	38
Mean	37.8	Mean	25.0	Mean	17.0
Group IV		**Group V**		**Group VI**	
10	9	4	−5	6	−23
12	23	8	−11	9	−42
14	32	26	15	16	−10
19	20	29	37	25	−22
22	19	30	−3	28	−10
Mean	20.6	Mean	6.6	Mean	−21.4

Behavior No. 9 was rated with less certainty than any other behavior. Its mean *SD* as given in Table II is .7096. As recalculated according to the above weighting, the mean *SD* is .6966—a negligible difference.

Behavior No. 15 was rated with the most certainty The weighted mean *SD* is .6642; the weighted mean *SD* is .6552, again a negligible difference

A third calculation was made, chosen at random. The mean unweighted *SD* for behavior No 1 is .7065; weighted, 7067—practically identical. It was therefore not thought worth while to recalculate the other behaviors, since the differences are obviously insignificant.

The failure of this factor of certainty to render more uniform the calculations may be explained by a further correlation The relation between certainty and uniformity (as measured by the mean *SD*) is − 0015 This means, of course, that raters differed as much among themselves when fairly certain as when uncertain The only conclusion to be drawn from the estimates of certainty, then, is that they tended to confirm the basis on which the six groups of behaviors were chosen. The experimenter was evidently fairly successful in choosing those behaviors which should be rated with the most and the least certainty. The only difficulty was that certainty and uniformity had little or nothing to do with each other

IV

Some light may be thrown on possible reasons for these results by again referring to the study of the first summer. The twenty-six behaviors then measured were chosen as being classifiable under some one of nine traits supposedly indicative of introversion-extroversion. Intercorrelations of all behaviors grouped under a given trait were calculated for both sets of data, *i.e*, ratings and objective records The mean intercorrelation of these one hundred twelve intra-trait behaviors, according to the ratings, was .493 ± .093, the mean of the same one hundred twelve intercorrelations, according to the objective records, was .141 ± .121 The conclusion may therefore be drawn that the halo effect, inevitable in the ratings, worked in such a way as to cause the rater to rate similarly logically related behaviors such as those classifiable under a single trait The close relation between the intra-trait behaviors which is evident in the ratings may, therefore, be presumed to spring from logical presuppositions in the minds of the raters, rather than from actual behaviors.

This hypothesis helps to account for the failure of the frequently observed behaviors to be rated more uniformly than those rarely observed, or those never observed at all There is, of course, one other possibility—that the frequently observed behaviors are, indeed, accurately rated and that, by inferring from familiar behaviors, the imagined situations are rated with equal accuracy To take this latter position, however, is equivalent to saying that behaviors may be accurately rated by guessing at them when they have never been observed Few psychologists will be willing to concede this It seems a truer statement to say that even with ample opportunity for observing, the ratings were little or no better than guesses.

The results of the study of degree of certainty tend to confirm this view. Raters agreed no more closely when fairly certain than when uncertain. The writer ventures a suggestion as to why this was true Assuming, as above, that the halo effect was at work almost equally in rating observed and unobserved behaviors (since the latter were as uniformly rated as the former), in the case of the former the raters were able to recall incidents supporting their "halo-ized" impression—conveniently forgetting, of course, the incidents which did not support this impression Such ratings were therefore marked certain In the case of the unobserved behaviors no such incidents could be called up, the impression remained, but could not be factually supported, and was therefore marked uncertain That the raters agreed to a considerable degree indicates that their halos coincided to some extent That they differed indicates that each rater had his own private halo.

The writer concludes, therefore, that under these conditions, apparently optimum for rating purposes, ratings on specific behaviors were so largely colored by the surrounding halo as to be quite invalid. Neither frequent observation of the behavior being rated, frequent recording of it, nor weighting the scores in accordance with the degree of the rater's certainty had an appreciable effect on uniformity of rating Behaviors never recorded, never seen, and those felt to be highly uncertain were as uniformly rated as those at the opposite extremes Since the guessed ratings are of highly questionable validity, must not the others, which are no more uniform, be almost equally invalid?

A METHOD FOR JUDGING THE DISCRIMINATION OF INDIVIDUAL QUESTIONS ON TRUE-FALSE EXAMINATIONS[1]

C. H WHELDEN, JR AND F J J DAVIES

Yale University

I INTRODUCTION

The following introductory comment on the scoring of true-false examinations is merely for the purpose of making clearer some of the later portions of this article It is generally recognized that whatever other method of scoring a true-false examination may be used, the method of scoring by the total of right answers is not satisfactory. If the number of questions is at all large, a man would be apt to get half his answers correct by sheer guessing.

A number of methods of scoring designed to eliminate the effects of guessing are possible, and the method adopted for any given case will depend upon the particular conditions of that case One such method is the following, which was adopted at the Yale Law School for its true-false examinations. Preliminary warning is given that guessing will be penalized, that if an answer has to be guessed it had better be omitted entirely. The examination is scored by adding the sum of omitted answers to twice the sum of wrong answers. The lowest score is then best and the highest is worst. The theory underlying the method is that guessing will be greatly minimized if not entirely eliminated, and that simple lack of information on a given question is not to be scored against so heavily as definitely wrong information on that question.

The minimizing of guessing through the preliminary warning seems to be accomplished. A brief investigation has given about as certain an indication of that result as is possible Out of seven examinations given in June, 1928 it was found on all but one that the group of men with the highest average law grades for the year (the top third of the group taking any one examination) had a smaller proportion of their examination scores accounted for by omitted answers than did the group of men with the lowest average law grades for the year (the bottom third of the group taking any one examination). On the sixth examination, as they are listed in Table I, the proportion ran slightly the other way but not sufficiently so to destroy the significance of the evidence given by the other six The indication might be taken to be that the best men guessed more than the poorest men. Logically the poorest men would be expected to do the most guessing, as anyone with much to gain and little to lose by taking a chance is most apt to take the chance The indication, therefore, is rather that guessing was so well minimized that the poorest men did not reveal their relatively greater propensity in that direction and that the best men, not realizing their own limitations so clearly as the poorest, gave evidence of incorrect information where the poorest men admitted lack of all information. (See Table I)

---

[1] The substance of a report prepared for the Yale Law School and the Department of Personnel Study of Yale University

One of the chief objects of a true-false examination and of its scoring, as in the case of any examination, is, of course, to discriminate adequately between men of different ability. A true-false examination correctly scored may give such discrimination fairly well, but at the

TABLE I.—THE PERCENTAGES OF TOTAL GROUP SCORES DUE TO WRONG ANSWERS AND TO OMITTED ANSWERS ON SEVEN TRUE-FALSE EXAMINATIONS, YALE LAW SCHOOL, JUNE, 1928

Course	Group *X*, top third of the men taking each examination—according to their law grades for year		Group *Z*, bottom third of men taking each examination—according to their law grades for year	
	Per cent of total score due to		Per cent of total score due to	
	Wrong answers	Omitted answers	Wrong answers	Omitted answers
I	88.3	11.7	84.7	15.3
II	94.2	5.8	90.9	9.1
III	93.8	6.7	92.9	8.1
IV	92.8	7.2	92.4	7.6
V	94.6	5.4	91.7	8.3
VI	96.0	4.0	96.6	3.4
VII	94.0	6.0	90.8	9.2
Seven examinations	93.4	6.6	91.3	8.7

NOTE —The table is designed for the comparison of Group *X* with Group *Z* If the table is used for a comparison of wrong answer percentage with omitted answer percentage within either group, it should be remembered that wrong answers are multiplied by two before going into the total score same time certain questions on the examination may give it much more adequately than others A question, for example, may contain some subtle ambiguity which will distract the better men but will not trouble the poorer men at all, another may be so difficult that only the exceptional man can answer it correctly while all others fail to comprehend it regardless of the varying degrees of their actual ability. Another question may be so simple or fundamental as to be answerable by practically anyone who has been exposed to the subject-matter

In order that true-false examinations may be improved in construction so as to discriminate more adequately between men, it is important that a method shall be available for judging the relative power of discrimination of the different questions which have been included in such examinations in the past. As a preliminary to the development of any such method it should be recognized that wherever possible the basis for the judgment on the individual questions should be the same as the basis selected as correct for the scoring of the examination. The discrimination given by the examination as a whole is only the net result of the discrimination given by the individual questions.

II. THEORY

If the relative abilities of men in the general field under examination can be determined independently of the specific examination which is to be judged, it can be said that the examination, if it discriminates correctly, should give those men scores which will vary in accordance with the independently determined measures of their respective abilities; such scores may be called Standard Examination Scores. For the sake of initial simplicity in exposition, in spite of the criticism already made of scoring by the total of right answers, consider the case of an examination of one hundred questions scored on the basis of total right answers, but with guessing supposed to be non-existent.

Suppose that the examination is so perfectly discriminating as to give a score of 80 to any man whose independently measured ability in the field is 80 on a scale of 100 Suppose there are one hundred such men. Each one will have given the right answer to eighty questions. If all the questions are of exactly equal caliber, not all the one hundred men will have given the right answer on any one question or the wrong answer on any one question. By definition the questions are of equal caliber In the hands of this group of men of equal ability each question will consequently be answered right by eighty men and wrong by twenty men, that is, the standard man-score per question will be 80 on the basis of right-answer scoring

Similarly, if there is another group of one hundred men, each of whom has an independently measured ability of 60 on a scale of 100, each one of the one hundred questions of equal caliber will be answered right by sixty men in this group. If the men in the first group obtained scores of 60 instead of 80 and the men in the second group scores of 45 instead of 60 (that is, if in the first group each question was answered right by sixty men and in the second group each question was answered right by forty-five men) the examination would have been one of greater difficulty, but each question, as well as the examination as a whole, would still be discriminating correctly between the men of the two grades of ability The men of the second group are still shown as possessing three-fourths as much ability as the men of the first group.

If one question, however, is answered right by eighty men in the first group and by eighty men in the second group, it does not discriminate between the two grades of ability. If all the questions were of the same caliber and gave this result, the score of each of the two hundred men on the examination would be 80 and the examination would not be discriminating as between the two groups of men of different ability. Similarly, if a question is answered right by eighty men in the first group and only forty men in the second, the question is over-discriminating. If a question is answered right by forty men in the first group and eighty in the second, the question gives reverse discrimination

Go back to the case of perfect discrimination by each question If there had been fifty instead of one hundred men in each of the two groups, the number getting each question right would obviously have been for each group one-half of the number found in the case of one hundred men to a group. If, on the other hand, there had been two hundred instead of one hundred questions on the examination, if all other conditions and assumptions remain the same, the examination score of an 80 man would have been 160, and the score of a 60 man, 120. If in this latter case there are also two hundred men in each group, each question will be answered right by one hundred sixty men of the first group and by one hundred twenty men of the second group.

Whatever the standard examination score is for a group of men of given ability as independently measured, the number of men in the group which will get right each one of the set of perfectly discriminating questions of average difficulty is given by the product of the standard score and the ratio of the number of men in the group to the number of questions on the examination. If the number of men of given ability in one group is fifty and the number of questions is two hundred, the standard score for the men of that group being, say, 160, the number of men getting each such perfect question right will be $(160 \times {}^{50}/_{200})$; that is, 40. This relationship must be used to reduce the standard question-score per man to the standard man-score per question, that is, to reduce the standard examination score to what is hereafter called the standard corrected examination score.

Not all the questions will be simply of average difficulty. One will be slightly more difficult, another slightly less, but they may still discriminate correctly between the grades of independently measured ability. In order to judge quickly the degree of discrimination shown by any question, it will be convenient to express the standard man-scores per question as ratios of one another. If, for example, the standard corrected score for a group of men of 80 ability is 30 and for a group of men of 60 ability is 15, the ratio of the score for the first group to the score for the second is 2.00. Then if on a given question twenty men of the first group and ten men of the second answer correctly, although the respective scores are not the standard corrected scores, it is at once clear that the question still discriminates correctly between the two groups, for the ratio of the scores made on the question is 2.00.

When the whole number of men taking an examination is divided into groups according to their independently measured abilities in the field (three such groups will generally be advisable for the purposes of the later judging of the questions), it will rarely, if ever, happen that all the men in any one such group will have the same independent measure of ability. The grades taken as measuring this general ability will vary within each group. The independent measure of ability for each group of men as a whole will be the average of their individual measures. Correlative with this average for the group there will be an average standard examination score in place of the single definite standard examination score implied by the discussion which has preceded.

The fact that the standard score for each group is in the nature of an average taken from variable measures requires another modification in the criteria used for judging the discrimination of the individual questions. Allowance must be made for the existence of variability within each group, for it introduces the probability of chance-fluctuations in any measures of the operations of the group The practically certain limits of variability within any group, which does not depart too far from normal in its distribution, is given by the range of three times the sigma (Standard Deviation) of the group above and below the arithmetic average of the group The sigma of the independent measures of ability for each group must be found, and then translated, in the form of a percentage of the mean from which it is calculated, to terms of standard examination score, just as the average independent measure of ability is to be found and translated; the process of this translation will be described in detail. The translated average and the translated 3-sigma range give a standard range of corrected examination scores for each group.

For example, a group of men with average independently measured ability of 80 may have a standard range of corrected examination scores of 60 to 40; and a group whose independent average is 60, a range of 30 to 20 Respective scores on a given question which fall within these ranges will now be the index of a correctly discriminating question The actual scores might be 60 and 20 respectively or 40 and 30 respectively. Either pair of scores must be taken as indicating correct discrimination since the two scores of each pair are within their respective 3-sigma limits of variation.

To keep the advantage of having ratios between scores, instead of a comparison between the scores themselves, for judging the degree of discrimination shown by a question, the limiting scores for the groups may be expressed as ratios of one another (the upper limit of one group as a ratio of the lower limit of another, and the lower limit of the first as a ratio of the upper limit of the second). The result is a range of ratios, instead of a single ratio, to be used as the criterion for discrimination. In the example last given, the range of ratios for criterion is $^{60}/_{20}$ to $^{40}/_{30}$, or 3.00 to 1 33

The same theory and method apply whatever the system used for scoring the true-false examination In the case of the Yale Law School examinations to which the method has been applied the system of scoring, as explained before, was twice the sum of wrong answers plus the sum of omitted answers. The transition from question-score per man to man-score per question follows here just as it does in the case, used purely for illustrative purposes, where scoring is by the total of right answers The question is always scored by the system which is used to score a man on the whole examination. In this case, therefore, each answer is scored for each of the groups of men by twice the sum of the men getting it wrong plus the sum of the men omitting it The theory no longer says, as in the case of the preceding illustrations, that of a group of men, all of 80 ability as independently measured, 160, say, if that is the standard corrected score for the group, will give the right answer on each of a set of questions of average difficulty and correct discrimination. It says instead that the men of such a group will on such a set of questions of equal caliber give the right answer to one question just as frequently as to any other, will give the wrong answer to one question just as frequently as to any other, and will omit one question just as frequently as any other. The frequency of answering right, answering wrong, and omitting to answer depends upon the level of ability of the men in the group The theory says, in other words, that the man-score per question in such a case will be the standard corrected examination score, all scoring being by a common system.

The one important point so far omitted from consideration is the matter of transition from the independently determined measures of ability in the field to the scores on a particular examination. In speaking of independently determined measures the meaning is simply that such measures must not be determined solely or primarily from the results of the examination of which the questions are to be judged for discrimination.

In the case of the Yale Law School true-false examinations the independent measure of ability of a man was simply his average law grade for the year. Such a grade was felt to be sufficiently independent of a man's showing on any one true-false examination to which the method of judging questions might be applied. The man's law grade for the year is the average of all his grades in all his courses Generally a man would take about five courses and in approximately three of these, as the general case, would be given a true-false examination. Almost without exception such a true-false examination would not stand alone in the course but would be supplementary to a written examination of the usual type. A man's showing on any one true-false examination in any one course is, therefore, but a small part of his average law grade for the year in all courses. Such average year grades are not, of course, precise measures of ability in the field of law Such grades, however, and it is the only important point in this connection, will be as good a reflection of the relative differences in ability of a group of men as is possible with anything so arbitrary in nature as an academic grade.

The connection between the independently determined measures of ability and the scoring of the given examination of which the questions are to be judged for discrimination may be made by a simple transmutation. A generally satisfactory method of making the transmutation for those scores which are necessary in the establishment of criteria for the judging of questions, without any waste of time over those scores which are not necessary for the purpose, may be illustrated by the procedure followed in the case of the Yale Law School examinations.

The average or year-grades of all law students for the year concerned are tabulated in a frequency series and cumulated downward from the higher grades On the examination to be judged the scores are similarly tabulated and cumulated downward from the lower scores, since on the system of scoring used a low score means a better showing than does a high score Suppose there are three hundred men represented in the general distribution of year-grades and ninety in the distribution of the scores on the particular examination The ninety men taking the examination have been divided into three groups of equal size according to their year-grades Suppose that one of these groups has an average of 80 in year-grades In the general distribution of year-grades there is found the percentage of the three hundred men with grades of 80 or higher In the distribution of examination scores this same percentage of the ninety men there included will have attained or exceeded a certain score (exceeded here in the sense of getting a lower score). This score is the transmuted equivalent of the eighty in year-grades. When this score is multiplied by the ratio of the number of men in the particular group (30 in this case) to the number of questions on the examination, the product is the average standard corrected examination score for the group The calculation of the 3-sigma range of variation follows directly at this point, in accordance with the method already indicated and explained more precisely in the section on Procedure

This method of transmutation assumes only that grades or scores on any one examination of any kind in the Law School will tend to follow in their actual significance a distribution which will be fairly uniform with the distribution of final year-grades given to the students in the school The assumption is one of similarity and not of coincidence So long as the particular examination is intended for no unusual purpose but simply for the customary testing of relative abilities, and so long as the number of cases in the distributions of grades and scores is fairly large, the assumption seems justified.

The rules for practical judging of the questions, as well as the exact method of establishing the criteria for judgment, are given in the section on Procedure. The language of this section, which follows shortly, applies specifically to the Yale Law School examinations but is easily generalized The actual rules find their premise in the underlying theory of the system, as that theory has been summarized here, but they have modified and developed out of practical experience in judging questions on seven true-false examinations given in the Yale Law School in June, 1928 No new modifications of these rules were found necessary when the system was applied to other examinations given in the school in February, 1928 and in February and June, 1929, but they should be considered as still open to modification as more experience with the system is gained. As given in the section on Procedure the rules are for the most part self-explanatory

A test has been made of the validity of the results obtained in the application of this system of judgment to true-false examinations given at the Yale Law School between 1928 and June, 1929. The test consisted in a series of linear correlations between average law grades and the true-false examination scores.

It should be remembered that the object of the system of judgment is to classify the questions according to the nature of the discrimination they tend to show as between students of different abilities. The standard for judging such discrimination by the questions is based on the discrimination actually made between students by their law grades. Thus, if the judgments given by the system are correct, the scores made on an examination composed entirely of *discriminating* and *over-discriminating* questions would have a high degree of positive correlation with the law grades of the men taking the examination; those made on an examination composed entirely of *non-discriminating* questions would have a very low correlation; those made on an examination composed entirely of *reversely discriminating* questions would have a negative correlation

The test, in its particular form, was suggested by Mr Paul W Burnham of the Department of Personnel Study at Yale The results of the test are given in Table II. It will be observed that throughout the eleven examinations listed in the test the correlations of law grades with the actual examination scores are lower than the correlations of law grades with the scores which would have been obtained if each examination had consisted only of its *discriminating* and *over-discriminating* questions, and that these latter correlations in themselves are positive and relatively high. It will be observed, further, that the correlations of law grades with the scores which would have been obtained if each examination had consisted only of its *non-discriminating* questions are in still another category of distinctly low degree, and that the correlations of law grades with scores based entirely on *reversely discriminating* questions are uniformly negative The degree of success in the application of the system obviously varies from examination to examination, but shows a marked variation downward in only one of the eleven examinations tested (number VI as they are given in Table II) The correlations obtained are such as apparently to establish the validity of the judgments.

TABLE II —CORRELATIONS BETWEEN AVERAGE LAW GRADES AND TRUE-FALSE EXAMINATION SCORES, FOR ELEVEN EXAMINATIONS GIVEN AT YALE LAW SCHOOL, 1928–1929

Examination	Coefficients of correlation			
	Law grades with actual examination scores	Law grades with examination scores that would have been obtained if examination had consisted entirely of its—		
		Discriminating and over-discriminating questions	Non-discriminating questions	Reversely discriminating questions
I	71	81	29	– 26
II	82	87	40	– 27
III	69	78	41	– 35
IV	66	69	29	– 24
V	75	84	29	– 36
VI	46	58	40	– 08
VII	57	72	10	– 31
VIII	50	75	12	– 63
IX	51	81	– 02	– 54
X	58	67	28	– 24
XI	56	71	36	– 48

NOTE —Coefficients are positive except where indicated as negative

III. PROCEDURE

The following are instructions for judging questions on true-false examinations of Yale Law School. Each numbered paragraph is hereinafter referred to as a section.

1. On each examination paper for a given examination place the student's law grade (average grade for year in all courses) as the independent measure of the student's ability.
2. For the given examination group make a distribution plot of the law grades, by tenths of per cent (Form 1)
3. From Form 1 divide the examination into three sub-groups according to law grade, each sub-group containing the same number of students Call these three groups, in descending order of law grade, X, Y, and Z, respectively

NOTE —Should the examination group not be exactly divisible by three, it will be necessary to omit one or two cases taken from the middle of the group, such cases to be given no further consideration.

4. According to the division determined in section 3, sort the examination papers into the three sub-groups X, Y, and Z.
5. Tabulate law grades for each sub-group X, Y, and Z, using class-interval of one per cent (Form 2).
6. Plot separately for each of the sub-groups X, Y, and Z the wrong answers and the omissions on each question in the examination (Form 3)
7. From Form 3 transfer to the respective columns X, Y, Z, and Total on Form 4 the scores of each question.

NOTE —The score is to be calculated by using the same basis as that used by the Law School for scoring a student's examination paper, *e g*, in June, 1928 the student's score was given by twice the sum of his wrong answers plus the sum of his omitted answers

8. Tabulate the average law grades last reported by the Law School for all students in the school, using a class-interval of one per cent Then cumulate downward from the high grades (Form 5)
9. Tabulate the scores on the given examination (sub-groups X, Y, and Z combined), using a class-interval of one point Then cumulate downward from the low scores (Form 6).
10. From Form 2 calculate the arithmetic mean and "sigma" of law grades for each sub-group X, Y, and Z.
11. For each sub-group X, Y, and Z use the following procedure
    - (*a*) Take the sub-group mean law grade as found in section 10 and find in the cumulative distribution column of Form 5 the number of students having that or a higher grade. Express this number as a percentage of the total number of students, *i e*, of N on Form 5
    - (*b*) Apply this percentage to the number of students in the examination under review, *i.e.*, N on Form 6; in the cumulative distribution of Form 6 locate the figure so gained and read the corresponding score in the examination score column of the same Form This score is the *average standard examination score* for the group

(c) Multiply this average standard examination score by the ratio

$$\frac{\text{Number of students in sub-group } (X,\ Y,\text{ or } Z)}{\text{Number of questions on examination}}$$

The result is the *average standard corrected examination score* for the group.

(d) Multiply this average standard corrected examination score by the ratio

$$\frac{\text{sigma (as found in section 10)}}{\text{mean (as found in section 10)}}$$

The result is the *standard corrected score value of the sigma-variation* for the group

(e) Three times this standard corrected score value of the sigma-variation (just found in *d*) added to and subtracted from the average standard corrected examination score (found in *c*) gives the *standard range of scores* for the group

12. List the standard range of scores for the three sub-groups on Form 4 and express them as ratios, thus:

$$\frac{X}{Y} \qquad \frac{\text{Upper limit of } X}{\text{Lower limit of } Y} \text{ to } \frac{\text{Lower limit of } X}{\text{Upper limit of } Y}$$

$$\frac{Y}{Z} \qquad \frac{\text{Upper limit of } Y}{\text{Lower limit of } Z} \text{ to } \frac{\text{Lower limit of } Y}{\text{Upper limit of } Z}$$

$$\frac{X}{Z} \qquad \frac{\text{Upper limit of } X}{\text{Lower limit of } Z} \text{ to } \frac{\text{Lower limit of } X}{\text{Upper limit of } Z}$$

13 *Grade as N* (see section 16 below), those questions which have in each sub-group X, Y, and Z scores less than the lower limit of the standard range of scores for sub-group X as found in section 11*e*

NOTE.—If a question is so easy that in each sub-group the score is less than would *normally be expected in even the best sub-group*, it is obviously a non-discriminating question If a question is difficult and in all sub-groups only a very few men get it right, it is advisable to apply the tests for discrimination given below, since it is more in the nature of a difficult question to be discriminating

14 For all remaining questions fill in the ratio columns on Form 4 as follows: Calculate and enter the X/Z ratios, calculate and enter either the X/Y ratios or the Y/Z ratios, whichever has the smaller upper limit in the Standard Ratio Ranges (as found in section 12 and listed on Form 4) In calculating these ratios consider a zero score as a score of one X/Y will normally be the second ratio used

NOTE —Two ratios for the questions are sufficient for grading purposes, the third ratio, when necessary, being determined by the relation between the other two, thus. Y/Z equals X/Z divided by X/Y, and X/Y equals X/Z divided by Y/Z. Assuming for the moment fairly normal distributions within each sub-group, the Y/Z standard ratio range will have a smaller upper limit than the X/Y standard ratio range if the mean law grade of sub-group Z is farther below the mean of sub-group Y than the mean of sub-group Y is below the mean of sub-group X That is, the Y/Z ratio will be used in preference to the X/Y ratio if the Z group shows greater divergence from the Y group than the Y group shows from the X group *If the sub-group distributions are highly abnormal so that the 3-sigma ranges overlap badly, the system will be applicable only after extensive modification, if at all*

15. Grade according to the following rules the remaining questions on Form 4, *i.e.*, those questions not graded *N* under section 13

(*a*) *Grade as N* (see section 16 below):

(1) Questions with *X/Z* ratio greater than 1, but the total score of the question is less than three times the lower limit of the Standard Range of Scores for sub-group *X* as found in section 11*c*

NOTE.—It is felt unwise to label a question Reversal instead of Non-discriminating (or Over-discriminating instead of simply Discriminating) unless the question is of such difficulty that the total score for all three sub-groups is at least equal to three times the lowest score normally expected in the best sub-group.

(2) Questions with *X/Z* ratio equal to 1, or between 1 and the upper limit of the *X/Z* Standard Ratio Range as found in section 12

(*b*) *Grade as R* (see section 16 below). Questions with *X/Z* ratio greater than 1, and the total score of the question is equal to or more than three times the lower limit of the standard range of scores for sub-group *X* as found in section 11*c*

(*c*) *Grade as O* (see section 16 below). Questions satisfying *all* the following requirements:

(1) *X/Z* ratio less than the lower limit of its standard range as found in section 12, and

(2) *X/Y* ratio (or *Y/Z* ratio, whichever is used as determined in section 14) less than the lower limit of its standard range as found in section 12, and

(3) The third ratio no greater than the upper limit of its standard range as found in section 12, and

(4) Total score for the question equal to or more than three times the lower limit of the standard range of scores for sub-group *X* as found in section 11*c*

(*d*) *Grade as D* (see section 16 below):

(1) Questions that satisfy requirement (1) for an *O* question (in paragraph *c* just above) but do not satisfy one or more of the other requirements for an *O* question

(2) Questions with the *X/Z* ratio within the limits of its standard range as found in section 12

16. Definitions of classes of judgment:

N. Questions that do not discriminate between sub-groups. Each sub-group tends to react similarly to such questions, making the questions of no value as a test of comparative ability between higher and lower sub-groups

R. Questions which discriminate inversely between sub-groups. In such cases students with the higher law grades obtain poorer results than do students with the lower law grades

O. Questions that discriminate between sub-groups to a greater extent than is justified by the relative levels of law grades

D. Questions that make normal discrimination between sub-groups

FORM 1.—DISTRIBUTION PLOT OF YEAR LAW GRADES FOR STUDENTS TAKING TRUE-FALSE EXAMINATION IN , , JUNE 1928

	Unit of law grade	Tenths of law grade										Total
		0	1	2	3	4	5	6	7	8	9	
(*X*)	84			1								1
	3	1				1						2
	2								1			1
	1		1									1
	80			1								1
	79					1						1
	8	1, 1		1	1, 1		1			1		7
	7		1		1			1				3
	6		1, 1			1, 1						4
	5	1	1		1		1					4
	4			1, 1								2
(*Y*)	3		1			1		1		1	1	5
	2	1		1	1, 1		1					5
	1	1, 1			1		1	1	1			6
	70		1, 1		1		(1)[1]	(1)[1]				5
	69	1	1	1, 1		1	1, 1, 1			1		9
	8		1, 1		1, 1	1		1	1			7
(*Z*)	7	1, 1		1, 1, 1	1		1, 1	1		1, 1		11
	6		1		1	1, 1	1					5
	5	1	1	1, 1				1			1	6
	4				1, 1	1, 1						4
	3		1				1, 1					3
	2				1				1			2
												95

N B —The solid lines mark the division into three groups *X*, *Y*, and *Z*, each containing 31 cases.

[1] The two cases at the middle of the distribution which are omitted from all further consideration, making the whole group exactly divisible into three sub-groups.

FORM 2.—GROUP DISTRIBUTIONS OF YEAR LAW GRADES FOR STUDENTS TAKING TRUE-FALSE EXAMINATION IN , , JUNE 1928

Group X		Group Y		Group Z	
Grade	Number of students	Grade	Number of students	Grade	Number of students
84	1	73	1	67	11
3	2	2	5	6	5
2	1	1	6	5	6
1	1	70	3	4	4
80	1	69	9	3	3
79	1	8	7	2	2
8	7		—		—
7	3	Total	31	Total	31
6	4				
5	4				
4	2				
3	4				
	—				
Total	31				

FORM 3.—CHECK LIST, BY QUESTIONS, OF NUMBER OF STUDENTS IN EACH SUBGROUP X, Y, AND Z ANSWERING WRONG OR OMITTING ANSWER, ON TRUE-FALSE EXAMINATION IN . . , JUNE 1928

Question number	Group X		Group Y		Group Z	
	Wrong answer	Omitted answer	Wrong answer	Omitted answer	Wrong answer	Omitted answer
1	0	0				
2	12	2				
3	14	3	(similarly)			
4	5	0				
5	11	7				
6	3	8				
Etc	Etc.					

FORM 4.—JUDGING CRITERIA, SCORES, RATIOS, AND QUESTION GRADES FOR TRUE-FALSE EXAMINATION IN . . . , JUNE 1928

Sub-group	Standard range of scores	Standard ratio ranges	Score criterion
X *Y* *Z*	10–13 16–19 21–24	*X/Y*, 81– 53 *Y/Z*, 90– 67 *X/Z*, 62– 42	For *N* Each subgroup score 9 or less. For *R* and *V*: Total score 30 or more

Question number	Score				Ratio		Question grade
	X	*Y*	*Z*	Total	*X/Y*	*X/Z*	
1	0	3	4	7			*N*
2	26	23	30	79	1 13	87	*N*
3	31	39	33	103	79	97	*N*
4	10	20	23	53	50	44	*D*
5	29	31	30	90	94	74	*N*
6	14	8	8	30	1 75	1 75	*R*
Etc					Etc.		

FORM 5.—SIMPLE AND CUMULATIVE DISTRIBUTIONS OF YEAR LAW GRADES FOR ALL STUDENTS IN THE SCHOOL, JUNE 1928
(These Figures Are Hypothetical)

Grade	Number of students	
	Simple distribution	Cumulative distribution
90	1	1
89	0	1
88	4	5
87	2	7
86	10	17
85	3	20
62	8	284
61	4	288
60	7	295
59	2	297
58	3	300
Total (*N*) . . .	300	

Applying section 11 Suppose mean law grade for a Group *X* is 85, it is found here that 6.7 per cent of the total men in the school had a law grade of 85 or better

FORM 6.—SIMPLE AND CUMULATIVE DISTRIBUTIONS OF EXAMINATION SCORES ON TRUE-FALSE EXAMINATION IN . . . . , JUNE 1928
(These Figures Are Hypothetical)

Score	Number of students	
	Simple distribution	Cumulative distribution
20	1	1
21	1	2
22	0	2
23	4	6
24	3	9
.	. . .	
.		
.		
55	4	90
56	0	90
57	1	91
58	2	93
Total (N) . . .	93	

Applying section 11: The figure determined for Group X from Form 5 was 6.7 per cent; 6.7 per cent of the total men (93) taking this particular examination is 6; it is seen that the examination score of 23 was attained or bettered by 6 men; 23 is the average standard examination score of Group X.

THE SIGMAS OF COMBINED DISTRIBUTIONS CALCULATED FROM SIGMAS, MEANS, AND FREQUENCIES OF COMPONENT DISTRIBUTIONS[1]

C R GARVEY

National Scholar in Child Development, University of Minnesota

It is sometimes necessary to compute for a series of measurements, not only the mean and standard deviation of the series as a whole, but also the means and sigmas of several fractional distributions, components of the total or general distribution The labor of calculation is thus doubled. If it is necessary to fractionate the data on more than one basis, *i e*, into more than one system of fractional distributions, the labor involved is multiplied by the number of such independent systems of fractionation plus 1 (the general system of parameters).

Obviously, this increase in labor can be kept down in the case of the means, by finding the smallest fractional means first and using them to compute weighted means for the others Where x = each measure and $\bar{x}$ = the mean of N such measures, the formula would be

$$\bar{x}_{1,2} = \frac{(\bar{x}_1)N_1 + (\bar{x}_2)N_2}{N_1 + N_2} \tag{1}$$

in which subscript 1,2 refers to a distribution composed of distributions 1 and 2. In practice, one operation is saved by using the form

$$\bar{x}_{1,2} = \frac{Sx_1 + Sx_2}{N_1 + N_2} \tag{2}$$

in which S indicates summation.

This procedure is familiar to everyone. The present purpose is to extend it to the calculation of the standard deviations The writer wishes to claim no priority for the following original formula, its essential features having been given in a similar one by Yule many years ago But since a survey of such frequently encountered texts as those of Kelley, Garrett, Thurstone, Fisher, Rugg, and Pearl indicates that the method is not in general use, it is thought worth while to present it here.

---

[1] The writer wishes to thank Dr Florence L Goodenough and Phillip J. Rulon for reading the manuscript, without imposing any responsibility upon either of them

Where $\Delta = x - \bar{x}$

$$\sigma = \sqrt{\frac{S\Delta^2}{N}}$$

When $d = x - a\bar{x}$ and $a\bar{x}$ = arbitrary mean

$$\sigma = \sqrt{\frac{Sd^2}{N} - c^2}$$

Now

$$c = \frac{Sd}{N}$$

then

$$\sigma = \sqrt{\frac{Sd^2}{N} - \left(\frac{Sd}{N}\right)^2}$$

When $a\bar{x} = 0$, then $d = x$, and

$$\sigma = \sqrt{\frac{Sx^2}{N} - \left(\frac{Sx}{N}\right)^2}$$

but

$$\frac{Sx}{N} = \bar{x}$$

so

$$\sigma = \sqrt{\frac{Sx^2}{N} - \bar{x}^2} \tag{3}$$

and

$$\sigma^2 = \frac{Sx^2}{N} - \bar{x}^2$$

This is essentially the formula given to his students by the late James Arthur Harris (published in 1910),[1] and the one pointed out by Beardsley Ruml in 1916 [2]

Transposing and assigning subscripts designating fractional distributions, we have

$$\frac{Sx_1^2}{N_1} - \bar{x}_1^2 = \sigma_1^2 \text{ and } \frac{Sx_2^2}{N_2} - \bar{x}_2^2 = \sigma_2^2 \tag{4}$$

---

[1] The Arithmetic of the Product Moment Method of Calculating the Coefficient of Correlation *American Naturalist*, Vol XLIV, 1910, pp. 693–699; especially 695f

[2] On the Computation of the Standard Deviation *Psychological Bulletin*, Vol XVIII, 1916, pp. 444–446.

By summation and from equations (1) and (2) we have

$$\frac{Sx_1^2 + Sx_2^2}{N_1 + N_2} - \bar{x}^2_{1,2} = \sigma^2_{1,2} \tag{5}$$

In machine calculation a table is set up, in which the rows are numbered to correspond to the subscripts in the formulæ, and which has the following eight column headings:

1	2	3	4	5	6	7	8
Sx	N	Sx^2	$\frac{Sx}{N} = \bar{x}$	$\bar{x}^2$	$\frac{Sx^2}{N}$	$\frac{Sx^2}{N} - \bar{x}^2 = \sigma^2$	σ

Divide the data into fractional distributions of the greatest common denominator, so that any desired combined distribution can be made up from these fractional distributions without further subdivision. Sum all the measures in the first small distribution and place the sum opposite subscript 1 in column 1 Place the number of measures in column 2 Square each measure (from a table), sum the squares, and place this sum in column 3 The rest of the table is self-explanatory, each subsequent entry being derived from previous entries Completion of a row of entries to column 8 fulfills equation (3). The mean and sigma of a combined distribution can be obtained by summing columns 1, 2, and 3, and calculating the entries for subsequent columns from these sums. Any simultaneous portion or portions of these first three columns can be summed, and thus any desired distribution can be combined from the appropriate fractional distributions. This fulfills equation (5) for any desired combination of subscripts

In reviewing literature or in other cases where Sx^2 is not given we need a formula using sigma instead From equation (4)

$$\frac{Sx_1^2}{N_1} - \bar{x}_1^2 = \sigma_1^2$$

transposing and multiplying by N_1 gives

$$Sx_1^2 = (\sigma_1^2 + \bar{x}_1^2)N_1$$

similarly for Sx_2^2. Then equation (5) becomes

$$\frac{(\sigma_1^2 + \bar{x}_1^2)N_1 + (\sigma_2^2 + \bar{x}_2^2)N_2}{N_1 + N_2} - \bar{x}^2_{1,2} = \sigma^2_{1,2} \tag{6}$$

Substituting the value of $\bar{x}_{1,2}$ from equation (1), and extending to include n fractional distributions, we have

$$\frac{(\sigma_1^2 + \bar{x}_1^2)N_1 + (\sigma_2^2 + \bar{x}_2^2)N_2 + \cdots (\sigma_n^2 + \bar{x}_n^2)N_n}{N_1 + N_2 + \cdots N_n} - \frac{(\bar{x}_1)N_1 + (\bar{x}_2)N_2 + \cdots (\bar{x}_n)N_n}{N_1 + N_2 + \cdots N_n} = \sigma^2{}_{1,2,} \cdots {}_n \quad (7)$$

This is comparable with Yule's equation (7),[1]

$$N \quad \sigma^2 = \Sigma(N_m \cdot \sigma_m^2) + \Sigma(N_m \cdot d_m^2),$$

except that here, equation (6), we use the fractional means themselves, whereas Yule uses the deviations d_1, d_2, of these means from the general mean $\bar{x}_{1,2,} \cdots {}_n$. Where these means are small they may as well be used directly. Where they are extremely large, the deviations are more easily used Experimenters should always report N along with $\bar{x}$ and sigma, so that their data can be included in a summary, and so that reviewers can actually review the author's work, instead of being limited to merely quoting the author's conclusions. Of course, if a reviewer combines measurements made by separate authors, he must examine the comparability of the conditions under which the separate sets of measurements were taken, just as he must in case of sets of data taken by the same author or by the reviewer himself

---

[1] "An Introduction to the Theory of Statistics," 9th ed , London, 1929, p 142.

A NOTE ON THE DEFINITION OF THE HARMONIC MEAN

EUGENE SHEN

Kwang Hua University, Shanghai

The writer has found that students completing an introductory course to statistical method often have a very hazy notion on the meaning of the harmonic mean. While they can make computations according to the formula, and even remember that its application has to do with the calculation of average rates, they seldom appreciate the problem in its comprehensive setting They often fail to see that the correct use of the formula depends upon the joint operation of two factors: The way in which events take place and the way in which records are made.

As the harmonic mean in educational psychology usually relates to the question of time and work, we shall take an illustration in this field. A group of students competing in addition may be required either to work during a uniform amount of time or to finish a uniform amount of work. Usually, records are made of the amount of work finished in the given time, or of the amount of time necessary for the given amount of work. The arithmetic mean in the two cases would correctly give the average work per unit time and the average time per unit of work, respectively. For purposes of comparison, we can use either one with the reciprocal of the other. The harmonic mean is not called for.

Sometimes, however, data are given in terms of time per unit work when they are derived from a constant amount of time, or on the other hand, in terms of work per unit time when they are derived from a constant amount of work. In both cases, the arithmetic mean would be incorrect, and the harmonic mean should be used instead [1]

The writer has proposed a definition of the harmonic mean as a special case of the weighted arithmetic mean where the weights are equal to the reciprocals of the measures [2] The true average height of man is not given by an unweighted arithmetic mean of the heights of different groups by race, residence, or political affiliation, because the groups vary in size and should be given different weights in proportion to the population in each group. Similarly the unweighted arithmetic mean of a series of rates for a uniform amount of work is incorrect because the varying rates do not operate for the same length of time, and needs to be weighted accordingly. Clearly the length of time for which each rate operated is proportional to its reciprocal, and therefore the correct mean is derived by weighting each rate according to its reciprocal This is precisely what the harmonic mean is.

---

[1] If in the foregoing we substitute price for rate, money for work, and commodity for time, we have a problem most frequently found in the field of business and economics

[2] "The Foundations of Experimental Psychology," edited by Carl Murchinson Clark University Press, 1929, p. 839.

It is the opinion of the writer that the proposed definition leads to a more systematic and more comprehensive conception of the harmonic mean. He has found it of valuable assistance in clarifying the mind of such students as are apt to be refractory to other methods of approach. Mathematically it is of course equivalent and reducible to the usual definition, as the reciprocal of the arithmetic mean of the reciprocals of the measures:

$$\mathrm{HM} = \frac{\Sigma\left(\frac{1}{X}X\right)}{\Sigma(1/X)} = \frac{\Sigma(1)}{\Sigma(1/X)} = \frac{N}{\Sigma(1/X)}.$$

NOTE ON THE STANDARD ERRORS OF THE STANDARD ERRORS OF ESTIMATE AND MEASUREMENT

CHESTER E KELLOGG AND KENNETH W SPENCE

McGill University

In the course of research on the reliability of the high-relief finger maze, recently completed by the junior author of this note, and to be published soon as part of a comprehensive study, occasion arose to compare the standard errors of measurement of intelligence tests and various methods of scoring maze records. In order to have a check on the validity of the conclusions drawn, we derived formulas for the standard errors of these and related measures.

It might naturally be supposed that the standard error of a standard error of measurement could be determined, as in the case of an ordinary standard deviation, by dividing by $(2n)^{1/2}$. In the case of the standard error of estimate, $\mathrm{SD}_{est} = \mathrm{SD}(1 - r^2)^{1/2}$, the corresponding formula does hold good. For taking differentials, we have.

$$d\mathrm{SD}_{est} = (1 - r^2)^{1/2} d\mathrm{SD} - \frac{\mathrm{SD}\, r\, dr}{(1 - r^2)^{1/2}}$$

Squaring, summing, and dividing through by n,

$$\mathrm{SD}^2_{sd_{est}} = (1 - r^2)\mathrm{SD}^2_{sd} + \frac{\mathrm{SD}^2 r^2 \mathrm{SD}_r^2}{(1 - r^2)} - 2r\mathrm{SD}\, r_{rsd}\mathrm{SD}_r\mathrm{SD}_{sd}$$

Assuming approximate normality, and using formulas 32*a*, 108*b*, and 125*a* from Kelley's "Statistical Method," this becomes:

$$\frac{(1 - r^2)\mathrm{SD}^2}{2n} + \frac{\mathrm{SD}^2 r^2(1 - r^2)^2}{(1 - r^2)n} - \frac{2r\mathrm{SD} \times r}{2^{1/2}} \times \frac{(1 - r^2)}{n^{1/2}} \times \frac{\mathrm{SD}}{(2n)^{1/2}},$$

which reduces to $(1 - r^2)SD^2/2n$

Accordingly,

$$\mathrm{SD}_{sd_{est}} = \frac{\mathrm{SD}(1 - r^2)^{1/2}}{(2n)^{1/2}} = \frac{\mathrm{SD}_{est}}{(2n)^{1/2}}. \quad \text{Q.E.D}$$

Similarly, for the standard error of measurement, $SD_{meas} = SD(1 - r)^{1/2}$, we have, taking differentials

$$d\mathrm{SD}_{meas} = (1 - r)^{1/2} d\mathrm{SD} - \frac{\mathrm{SD}\, dr}{2(1 - r)^{1/2}}$$

Squaring, summing, and dividing through by n,

$$\mathrm{SD}^2_{sd_{meas}} = (1 - r)\mathrm{SD}^2_{sd} + \frac{\mathrm{SD}^2\mathrm{SD}_r^2}{4(1 - r)} - \left(2\mathrm{SD}(1 - r)^{1/2} \times r_{r_{sd}} \times \mathrm{SD}_r \times \frac{\mathrm{SD}_{sd}}{2(1 - r)^{1/2}}\right)$$

$$= \frac{(1 - r)\mathrm{SD}^2}{2n} + \frac{\mathrm{SD}^2(1 - r^2)^2}{4n(1 - r)} - \left(\frac{\mathrm{SD} \times r}{2^{1/2}} \times \frac{(1 - r^2)}{n^{1/2}} \times \frac{\mathrm{SD}}{(2n)_{1/2}}\right)$$

which reduces to $\mathrm{SD}^2(1 - r)(3 - r^2)/4n$

Accordingly,

$$\mathrm{SD}_{sd_{meas}} = \mathrm{SD}(1 - r)^{1/2}\frac{(3 - r^2)^{1/2}}{(4n)^{1/2}}$$

$$= \frac{\mathrm{SD}_{meas.}(3 - r^2)^{1/2}}{(4n)^{1/2}}$$

To facilitate calculation of the $\mathrm{SD}_{sd_{meas}}$, we have tabulated representative values of $(3 - r^2)^{1/2}/(4n)^{1/2}$

n	r							
	00	10	40	65	80	90	95	99
25	.1732	1729	1685	1605	1536	.1480	1448	1421
30	.1581	1578	1538	.1466	1402	1351	1322	.1297
35	1464	1461	1424	1357	1298	.1250	1224	1201
40	1369	1367	1332	1269	1214	.1170	1145	1123
45	1291	1289	1256	.1197	1145	.1103	1080	.1059
50	1225	1223	1191	.1135	1086	.1046	.1024	.1005
55	1168	1166	1136	1082	1036	0998	0977	.0958
60	1118	1116	.1088	1036	0991	.0955	0935	0917
65	.1074	1072	1045	0996	0953	0918	0898	.0881
70	1035	.1033	1007	0959	0918	0884	0866	.0849
75	1000	0998	0973	0927	0887	0854	.0836	0821
85	0939	0938	0914	0871	0833	0802	0785	0771
100	0866	0865	0843	0803	0768	0740	0724	0710
150	0707	.0706	0689	0655	0627	0604	0591	0580
200	.0612	0611	0596	0568	0543	0523	0512	.0502

Although it is not likely to be much in demand at present, we have also derived a formula for the standard error of the standard error of estimate of true score, $\mathrm{SD}_{\infty.1} = \mathrm{SD}_1(r_{11} - r^2_{11})^{1/2}$. (Kelley, formula 160.)

$$d\mathrm{SD}_{\infty.1} = \frac{\mathrm{SD}(dr - 2rdr)}{2(r - r^2)^{1/2}} + (r - r^2)^{1/2}d\mathrm{SD}.$$

Squaring, summing, and dividing through by n,

$$\mathrm{SD}^2_{sd_{\infty 1}} = \mathrm{SD}^2\mathrm{SD}^2_r\frac{(1 - 2r)^2}{4(r - r^2)} + (r - r^2)\mathrm{SD}^2_{sd}$$

$$+ 2\mathrm{SD}(r - r^2)^{1/2}(r_{rsd}\mathrm{SD}_r\mathrm{SD}_{sd} - \frac{2r\ r_{rsd}\mathrm{SD}_r\mathrm{SD}_{sd})}{2(r - r^2)^{1/2}}$$

$$= \mathrm{SD}^2\frac{(1 - 4r + 5r^2 + r^3 - 4r^4 - r^5 + 2r^6)}{4n(r - r^2)}.$$

Accordingly,

$$\mathrm{SD}_{sd_{\infty 1}} = \mathrm{SD}(r - r^2)^{1/2}\frac{(1 - 3r + 2r^2 + 3r^3 - r^4 - 2r^5)^{1/2}}{(4nr(r - r^2))^{1/2}},$$

$$= \mathrm{SD}_{\infty 1}\frac{(1 - 3r + 2r^2 + 3r^3 - r^4 - 2r^5)^{1/2}}{(4nr(r - r^2))^{1/2}}.$$

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

CONDUCTED BY FRANCES M. FOSTER

Educational Psychology, by Monroe, DeVoss and Reagan New York. Doubleday Doran, 1930. Pp XIII + 607.

Problems in Educational Psychology, by Gifford and Shorts. New York: Doubleday Doran, 1930 Pp XIV + 728.

These volumes are two from the excellently planned Teacher Training Series, edited by W. S. Monroe, one of the authors of the *Educational Psychology* The second text listed above was designed to supply companion readings to any basal text in educational psychology, but the authors acknowledge that they had the Monroe, DeVoss and Reagan text principally in mind when they planned their work.

The *Educational Psychology* is a well-constructed book exhibiting many excellent features. In the selection of the subject-matter a wise catholicity has been shown. The topics included are—The Physical Mechanism; Human Responses to Stimuli, The Learning Process; Learning in School Activities; Transfer of Training; Intelligence and its Measurement; Measurement of Achievement, Individual Differences, Characteristics of Children at Different Pedagogical Levels, The Psychology of Elementary-School Subjects, The Psychology of High-School Subjects; Mental Hygiene; and How to Study Pupils. Such a list can be made strong meat and in their efforts to keep the text at an elementary level a few of the topics necessarily become a little obscure. The "limit of improvement" is an illustration of this. Apparently the authors take up the position of believing in limits for manual habits only; the acquisition of knowledge may proceed with ever-increasing facility Which, of course, is a misinterpretation The authors have also hunted with the hounds and run with the hares in respect to subjective and objective observations although, as scientists, they have emphasized throughout their volume the data obtained from carefully controlled experiments The best feature of the book is the consistency with which technical terms have been used. When, for example, they wish to speak of intelligence derived from the use of a test as distinguished from intelligence as a theoretical concept they use "intelligence as measured" and thus get rid of any ambiguity. The Learning Exercises for the reader at the end of each chapter are truly exercises to promote further learning and are commendable.

The *Problems in Educational Psychology* is an anthology of excerpts or readings from works of 192 authors Thorndike and Woodworth are quoted the most frequently, and justly so. The selections appear to have been made more carefully than those of previous compilers such as Skinner, Gast and Skinner. By listing in each chapter "Suggested Problems" and "Supplementary Learning Exercises" some attempt to justify the selected title has been made.

Both volumes are well printed and strongly bound, and both are remarkably free from typographical errors. The reviewer wishes them the success they deserve. P. SANDIFORD.

University of Toronto.

Minnesota Mechanical Ability Tests, by D. G. Paterson, R. M. Elliott, L. D Anderson, H A. Toops, and E. Heidbreder Minneapolis: University of Minnesota Press, 1930. Pp. XXII + 526.

It is a genuine pleasure to study the methods—so thorough, so cautious, and often inventive—which characterize this excellent and important investigation. Within recent years there have been published few other studies of similar thoroughness, so insistent on the accuracy of the instruments of measurement, so critical of their validity There is reason to think that with this type of investigation (one would like to name those few others of equally fine caliber!) the measurement of human behavior has, within the last few years, achieved one more significant step in its advance towards recognition as an exact science In the field of mechanical ability undoubtedly this particular work is fundamental.

The authors set out to investigate the field of mechanical ability to determine adequate answers to the two main questions (1) Is "mechanical ability" one ability or many? and (2) How is it related to other traits such as verbal intelligence and motor ability?

By force of circumstances, similar to those experienced by investigators in the field of measurement of verbal intelligence, it became necessary to define mechanical ability as "that which enables a person to succeed in a definitely restricted range of vocational and trade school courses"

A survey of literature in the field yielded twenty-four tests relating to mechanical ability These were administered in a preliminary investigation to two hundred seventeen boys, all taking shop-work, in Grades VII and VIII in a junior high school. Nearly all of the tests were readministered after a suitable period of time, thus yielding a measure of reliability.

The criterion selected was shop grades, determined as objectively as possible in terms of quality of work, quantity of work, and information displayed in examinations.

The scores on every test were then correlated with (1) the criterion scores and (2) the average scores on two verbal intelligence tests Seven tests, yielding the highest correlation coefficients with the combined criterion, were selected to comprize the final battery

$$(R_{\text{intell batt}} = 07; R_{\text{crit batt}} = 59, R_{\text{crit+intell batt}} = 61).$$

The reliability coefficients of these seven tests ranged, however, from 65 to .80. By lengthening some, and imposing a time limit on others, the coefficients were raised, theoretically at least, to range from .80 to .93, expectations closely approximated in the final experiment.

In the experiment proper this battery (now known as the Minnesota Mechanical Ability Tests) was administered to one-hundred fifty incoming boys in the same grades at the same school. In addition, thirty-six other measures were determined, covering academic success, previous mechanical experience, interests, motor ability, anthropometric measures, social and economic status, and home influences

The validity of the criterion was determined on the basis of objective standards of judgement—the information factor on the basis of objective tests, and the quantitative factor on the basis of production. The careful construction of objective rating scales yielded a quality criterion (3:3 to 6.6 judges, depending upon the function measured) of reliability over .90 *The reliabilities of the different shop criterion approximated* .80. The correlations of the individual tests with the quality-quantity criterion, however, were so low that the quantity criterion was abandoned. *The validity correlation between battery and quality criterion alone was* .65 The validity of the battery in respect of each individual type of shop work was found to be about as good as that of most standard intelligence tests in respect of success in individual academic subjects

Analysis of results in accordance with the principles of Spearman suggests that specific factors rather than a single general factor characterize mechanical ability There is, however, some evidence of the presence of group factors Four-factors, on the other hand, were found to be unique namely intelligence, height, agility, and mechanical ability.

The authors thus conclude in the main (1) that mechanical ability as here defined is a unique trait, (2) consisting of a number of specific traits having possibly group factors in common.

The reviewer would have felt entirely satisfied had the three following questions been answered (1) During the actual testing process what control was there over the possibilities of leakage of information concerning the tests?

(2) In so far as the criterion fails to measure "originality" does it not fall short of what, with occupational reference, might be termed mechanical ability? (3) To what extent are these findings in agreement with those of a recent investigation into "mechanical aptitude" by John W. Cox, one of Spearman's pupils?

Comment on the valuable investigation here reviewed cannot be allowed to pass without reference to the excellence of the binding, printing, and general layout of the volume O L. HARVEY

University of Texas.

Problems of Science Teaching at the College Level, by Archer Willis Hurd. Minneapolis, Minnesota · University of Minnesota Press.

In these days of educational innovations and unproved panaceas for college ailments a bit of cautious, scientific investigation in the field of higher learning is like fresh water to the thirsty For some years college men have been protagonists for educational experimentation and research in the elementary and secondary schools. Student achievement and that relating thereto has been measured and probed. Some of the most loudly advocated new methods of teaching have been partially examined However, perhaps because the beam seemed bigger in the other fellow's eye, there has been comparatively little scientific investigation of teaching problems in college. It seems quite fitting that one should arise from the examined group and now lead some study of a similar nature in the realm of those who first made up the game

The work reported concerns itself with the following problems: (1) What differences in individual achievement in the study of anatomy is produced between those who work in groups of two and those who work in groups of four on a cadaver? (2) What is the effect of limiting the time given to laboratory work in human physiology or of partially replacing the laboratory work with library work? (3) What is the effect of eliminating laboratory work in the study of "Mechanics"? (4) What effects has class size on individual achievement in the physics courses of "Heat" and "Electricity and Magnetism"? (5) What influence does a high school course in physics have on the achievement of students in college physics?

The details of the conclusions drawn from the studies are, naturally, of special interest to college teachers of anatomy, physiology, and physics. The main finding likely to be of general interest is that class size in the courses investigated appears to have no influence on student achievement. In the words of the author, "Achievement seems to be more a matter of individual incentive, capacity, and effort."

Dr. Hurd has nicely evaluated his work when he says, "The studies . find their greatest value in actual, concrete illustrations of techniques in educational experimentation. They represent attempts to settle problems of teaching by methods of experiment used in science." The discussion of the outstanding references in the excellent bibliography and the section of the conclusions on suggested techniques for this type of experimentation strike the reviewer as being particularly noteworthy LEONARD B WHEAT

Teachers College, Columbia University.

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII May, 1931 Number 5

THE ASSOCIATION FACTOR IN INTELLIGENCE TESTING*

S TOLANSKY

Fellow of Armstrong College, Newcastle-upon-Tyne

INTRODUCTION

A good deal of work has been done on the reliability of intelligence tests and the constancy of IQ but very little on the analysis of response error Holzinger[1] considers a response error (δ) as due to fluctuations in effort, emotional status, concentration, etc. He finds that this is roughly normally distributed. Stenquist[2] observes that, if the Terman tests are repeated on a group, seven per cent are twenty points out, and this is probably due to response error. An account is given in this paper of an investigation on the association factor in intelligence tests. The tests selected were from the American Army α Test, admittedly somewhat old Instead of allowing, say, the usual two minutes for a test, a modified technique was employed. The test was covered up The first question uncovered and the time taken to answer it was noted with a stop watch reading to one-tenth second The next question was uncovered and the time taken This procedure was repeated to the end of the test. Very little time was spent in covering and uncovering and the timing errors cancel out. Only those questions done in the time stipulated in the test are considered for the score By graphing times against the number of the question it is strikingly shown that certain questions (different as a rule for each person) give abnormal times There are three possible causes of delayed response to a question. They are. (1) Ignorance of facts;

---

* I wish to thank Mr Vernon Brown for assistance in the preparation of this paper

(2) inability or partial inability to answer, through weakness in reasoning; (3) delayed reaction due to an emotional association.

In a good intelligence test (1) is reduced to a minimum and the cause (2), which is that concerned with intelligence, becomes important However (3) cannot be neglected by any means Jung[3] has shown in his word association experiments, that words involving an emotional factor take a longer time for a response than do words which do not arouse such a factor I have found as a result of experience that a delay up to ten seconds in ordinary word association is sufficiently common to be considered normal Suppose then that in a test, which has not been completed in the stipulated time, there are words or ideas which call up associations emotionally toned, then, irrespective of intelligence, delays will take place over these. The net result is to reduce the IQ, for had the time not been uselessly absorbed more questions could have been answered There are, of course, associations to every question. These can be classified into the useful, helping to solve the problem, and the useless, of the Freudian or Jung type. If all the questions have been completed in the time, then the delays do not affect the number attempted The possibility of useless association is purely accidental, depending on previous experience, and this may partly account for disagreements in reliability and IQ constancy observations

EXPERIMENTAL METHOD

Two tests were selected, a true-false test and a multiple choice test. Both were given to five women and five men college students The reaction times were taken and the testees were carefully watched The questions taking abnormally long time were selected and the testee's attention drawn to them In some the delay was easily explained The testees immediately admitted lack of knowledge, or actual difficulty. The others, which predominate, delay was due to useless association The questions were dealt with individually and a regular psychoanalysis carried out In every case it was found that the cause of the delay was due to associations of emotionally toned complexes, some going back many years, some sexual, and some being repressed incidents of an unpleasant nature. In practically every case the examinee was hardly aware that the time taken was long, yet in some instances a change from an average of six seconds to a delayed time of seventeen seconds took place. It was often easy to tell from facial expression that the emotions were disturbed

It has been considered advisable to reproduce the two tests since frequent reference will be made to them To illustrate the effect of useless association two tests have been selected at random from the twenty and the detailed analysis given The results of the twenty tests are shown in Tables I and III and conclusions drawn from them are recorded

Description of Tests

TEST I

Type, multiple choice
Sixteen questions are to be answered
Time allowed—ninety seconds.
Subtest

1 It is wiser to put money aside and not spend it all so that you may: (*a*) Prepare for old age and sickness; (*b*) collect all the different kinds of money; (*c*) gamble if you wish

2. Shoes are made of leather because· (*a*) It is tanned, (*b*) it is tough, pliable and warm, (*c*) it can be blackened

3. Why do soldiers wear wrist watches rather than pocket watches? Because (*a*) They keep better time, (*b*) they are harder to break; (*c*) they are handier.

4 The main reason why stone is used for building purposes is because. (*a*) It makes a good appearance; (*b*) it is strong and lasting; (*c*) it is heavy

5 Why is beef better food than cabbage? Because: (*a*) It tastes better; (*b*) it is more nourishing, (*c*) it is harder to obtain

6 If someone does you a favour what should you do? (*a*) Try to forget it; (*b*) steal for him if he asks you to, (*c*) return the favour

7 If you do not get a letter from home which you know was written, it may be because (*a*) It was lost in the mails, (*b*) you forgot to tell your people to write, (*c*) the postal service has been discontinued

8 The main thing farmers do is to (*a*) Supply luxuries, (*b*) make work for the unemployed, (*c*) feed the nation

9 If a man who can't swim should fall into a river he should (*a*) Yell for help and try to scramble out, (*b*) dive to the bottom and crawl out, (*c*) lie on his back and float

10 Glass insulators are used to fasten telegraph wires because (*a*) The glass keeps the pole from being burned, (*b*) the glass keeps the current from escaping, (*c*) the glass is cheap and attractive

11. If your load of coal gets stuck in the mud what should you do? (*a*) Leave it there; (*b*) get more horses or men to pull it out, (*c*) throw off the load

12. Why are criminals locked up? (*a*) To protect society, (*b*) to get even with them, (*c*) to make them work

13 Why should a married man have his life insured? Because (*a*) Death may come at any time; (*b*) insurance companies are usually honest, (*c*) his family will not then suffer if he dies

14 In leap years February has twenty-nine days because (*a*) February is a short month, (*b*) Some people are born on February 29, (*c*) otherwise the calendar would not come out right.

15. If you are held up and robbed in a strange city you should (*a*) Apply to the police for help, (*b*) ask the first man you meet for money, (*c*) borrow some money at a bank.
16 Why should we have Congressmen? Because (*a*) The people must be ruled; (*b*) it ensures truly representative government, (*c*) the people are too many to meet and make their laws

EXPERIMENTAL RESULTS

In a test of this nature a variation in time for each subtest is expected because of the varying times required for reading, etc. This, however, is systematic and affects each person to a roughly equal degree Different rates of reading in different persons will seriously affect the result, which is a bad fault in the test. The times taken by the ten people *A* to *J* are given in Table I The subtests completed in ninety seconds by each individual are shown above the black line.

TABLE I

(Time Allowed—Ninety Seconds)

Problem	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
1	10 0*	7.2	9 0	5 2	5 4	8 8	15 0	4 8	13 0	15 1
2	6 2	4 0	3 7	4 0	3 8	3 8	7 2	2 0	6 8	4.0
3	7 0	4 3	3 0	5 0	5 6	6 3	12 5*	3 0*	5 9	6.3
4	5 6	4 2	3 9	4 0	6.0	7 0	7 6	2 5	4 2	4 6
5	4.7	6 7	3 9	2 0	5 8	7 0	6 0	2 5	4.1	3 3
6	5.0	6 0	3 9	3 0	3 4	7 4	5.0	3.0	5 0	6 0
7	9.2*	9 8*	6 2*	13 5*	7 4	8 2*	17 4*	4 2*	8 0*	6 0
8	4 8	3 0	3 0	3 0	4 0	3 9	6 1	3 2	5 6	6 0
9	6 2	6 5	5.2	5 0	7 0*	6 0	6 4	3 8	6 4	7 0
10	9 0	6 4	8 8	4 0	6 0	6 5	8 4	3 8	7 0	8 4*
11	5.4	6.2	5 0	12 0*	4.4	14 0*	10 0*	4.0*	17 2*	6 4
12	4 2	3 8	4 1	3 0	3 5	5.0	6 0	3 5	4 4	6.0
13	7 8*	8 0*	6 8*	5 0	6 5	7 5	7 8	3 8	5 2	8 0*
14	11 6*	5 2	8 0	4 2	4 5	6 5	15 0	4 0	7 4	6 0
15	6 0	5 2	5 0	5 0	4 6	5 0	8 0	2 8	5 0	6 8
16	17 2*	7 8*	5 0	9 0*	18 5*	11 0*	12 0*	10 0	15 0*	7 4*
Total time in seconds .	120	94	85	87	96	114	150	61	108	107
Mark score obtained .	11	13	15	13	15	10	8	14	13	11

The total times are given to the nearest second and below this the score obtained A full analysis will follow later The asterisk indicates useless association, the other long times are accounted for under headings (1) and (2) at the beginning of this article

The full analyses for *A* will now be given. *A* is a woman graduate twenty-two years old. It is convenient to classify the first part of each question as the statement and the three choices as I, II, III

1 Problem (1) takes 10 seconds Statement causes the following associations "Foresight—insurance—doctor—illness—child—insurance." *A's* sister is ill and the doctor called today He is the doctor of a lady friend who married an ailing husband. The husband died recently and the doctor had asked how the wife existed and was told by *A* that the husband was heavily insured (see question) A child was born to the lady during the last illness of the father *A* was surprised and cynical for the lady claimed she had no sex knowledge and confided in *A* that she did not sleep with her husband, a point which struck *A* very forcibly.

2 Problem (7) takes 9.2 seconds and choice II gave "lack of letters—urging at home—disappointment with mother—upbraiding" Two years ago *A* was in France for some months and her mother hardly ever wrote her although her sisters did. She was disappointed and complained on returning She often blames her mother for failure to write.

3 Problem (13) takes 7 8 seconds, the word "insured" giving, "insurance-agent—wife—peroxide—Peter." There was a long pause after this *A* said "Oh dash!" when the word "insurance-agent" slipped out That day her mother told her she had seen an insurance-agent who had married a fair-haired girl and now had a child, Peter Some years ago *A* thought this man keen on her (note the depreciating, perhaps jealous, remark of "peroxide" referring to the wife's fair hair) *A* said she was pleased he was now settled as he is a bit of a philanderer She was a little too anxious in disclaiming anything but friendly regard

4 Problem (14) takes a long time, 11 6 seconds, choice III gives two distinct trains of association

1 "Reform of Calendar—Julius Caesar—month of August (named after Augustus)—Harvest Festival in August" *A* stopped and related the following. She remembered when three years old falling off a seat in church during a harvest festival Someone picked her up and threatened to steal her from her parents She even remembers the dress she was wearing at the time

2. "Change in Calendar—St. Cuthbert's Church—vicar—scurvy trick" At the above church *A* had taken a prominent part in social life On removing she had neglected to say farewell to the vicar and thought this very bad The vicar was very conservative and would certainly object to the recent proposal for fixing the date of Easter

5 The last question took an abnormal time—17 seconds. Choice II caused a whole group of associations connected with the fact that *A* believes present day government is entirely undemocratic

This concludes the analysis. Altogether thirteen questions were done in time. This is indicated by the line at thirteen in Table I. Table II, *A* gives a detailed result for testee *A*. Useless association is indicated by the figure 1 in Tables II and IV, and where the time is long, due to the absence of useful association only, this is shown by a figure 2.* When the question has been wrongly answered a period is given Thus 1. indicates a wrong answer and useless association. *A* thought problem (10) was of a technical nature, was afraid, and guessed, wrongly Eleven marks were scored. The average normal time for questions, other than those marked by a 1, is five seconds. The five type 1 associations taking the long time of sixty-four seconds, *i.e.*, thirty-nine above the average Since thirty seconds more than the time allowed was taken, it is obvious that but for the useless association *A* would have completed the whole test, and obtained three more marks. Thirty-one per cent of the questions which are emotion-

TABLE II

Problem	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	Totals		
											1	2	
1	1										1		
2													
3						2	1	2.			1	2	3
4													
5													
6													
7	1	1	1	1		1	1.	1	1		8	.	2
8													
9					1						1		7
10	2	.	2							1	1	2	2
11		.		1		1	1	1	1		5		
12													
13	1	1.	1			2				1	4	1	2
14	1	.	2	.			2	2	2		1	4	1
15													
16	1	1		1	1	1	2	1	1	1	8	1	4
Total 1	5	3	2	3	2	3	3	4	3	3			
Total 2	2	2	2			2	2	1	1				

---

* Referred to as types 1 and 2 later

ally toned take seventy-one per cent of the time allowed From the analysis given there is obviously no doubting the strength of the emotional tone and its real effect. As seen from the table, *A* is not the most extreme case.

The second test, which was given to the same ten people, is as follows It is the true-false type. Sentences had to be arranged into sense and marked true or false There are twenty-four questions and one hundred twenty seconds are allowed. A mark is deducted for each error, as usual.

TEST II

1 Lions strong are
2 Houses people in live
3 Days there in a week eight are
4 Leg flies one have only
5 Months coldest are summer the
6 Gotten sea water sugar is from.
7 Honey bees flowers gather the from
8. And eat good gold silver to are
9 President Columbus first the was America of
10 Making is bread valuable wheat for
11 Water and made are butter from cheese.
12 Sides every has four triangle.
13. Every times makes mistakes person at
14 Many toes fingers as men as have
15. Not eat gunpowder to good is
16 Ninety canal ago built Panama years was the
17 Live dangerous is near a volcano to it
18 Clothing worthless are and for wool cotton
19 As sheets are napkins used never
20 People trusted intemperate be always can
21 Employ debaters irony never
22 Certain some death of mean kinds sickness
23 Envy bad malice traits and are.
24 Repeated call for courtesies associations

Full results are given in Tables III and IV Full analysis is given for *C* who has the best mark and finished in one hundred seven seconds. As usual, a mark is subtracted from the total for every error, to correct for guessing *C* is a science student (man) aged twenty.

---

1 Problem (10) takes 5 4 seconds Associations were "cornfield—fresh air—house—removal" Five years ago *C* had to leave his house, the family going to a country house next to *what was then* a cornfield He is now of the opinion that the

change was an advantage, but was doubtful then. He recalled his mixed feelings at going to live in the country.

2. Problem (19) takes 5.4 seconds, giving "napkin—set table—Christmas—spilled wine." Some years ago he had spilled a glass of wine with his napkin at a Christmas dinner at home. Christmas is the only occasion on which he gets wine. It was not replaced, he thought it very wasteful, ruined a tablecloth, and the whole family laughed at him. He vividly recalls the incident. The mental disturbance in the test was strong since he also answered wrongly, much to his surprise, on later reading his answers.

TABLE III

(Time Allowed—One Hundred Twenty Seconds)

Problem	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
1	5.0	2.4	3.0	1.7	3.0	2.0	4.0	1.9	2.4	4.1
2	3.7	2.4	1.9	2.2	1.7	2.0	3.0	1.0	3.0	2.2
3	4.3	2.0	2.2	1.4	2.0	2.4	3.3	2.4*	4.5*	2.2
4	4.3	2.2	2.2	2.4	2.1	2.2	4.8*	2.0	4.0	4.0
5	4.0	2.4	2.2	1.4	2.1	3.0	2.8	2.2	3.5	3.5
6	3.1	2.8	2.2	1.9	2.0	2.8	3.0	2.5	5.0*	2.2
7	3.3	2.4	3.0	2.2	2.0	3.2	3.0	2.5	4.7	2.8
8	12.0*	2.5	3.2	2.0	3.1	5.2*	2.4	4.5*	10.0*	2.6
9	2.6	3.0	2.8	1.6	4.0	2.5	2.8	2.0	3.0	4.0
10	3.5	1.4	5.4*	1.0	9.0*	2.2	2.6	2.2	3.0	3.0
11	3.8	10.5	4.0	2.0	6.0	3.9*	3.0	5.0	10.2*	2.1
12	3.2	1.9	2.6	1.6	1.0	2.5	2.8	2.0	2.1	2.8
13	12.5	3.7	2.7	9.0*	2.0	2.6*	3.0	4.7	3.9	4.0
14	3.6	3.1	5.4	2.4	3.0	5.0*	6.4*	6.0*	4.0	3.0
15	3.1	1.6	2.0	1.4	2.4	2.0	2.9	2.7	3.8	2.0
16	6.2	3.0	12.0	3.1	2.4	3.8	9.4	4.8	5.6	3.1
17	2.0	1.0	2.2	1.4	1.0	2.2	2.9	4.0	2.4	2.9
18	5.0	2.0	4.0	1.6	3.0	2.8	11.4*	3.0	4.0	3.2
19	16.4*	4.4*	5.4*	3.0*	2.6	3.0*	5.9	4.0	4.0	6.0*
20	6.4*	2.9	6.0	2.4	2.6	5.0*	7.4*	3.6	3.0	3.8
21	6.7*	4.0*	8.8	2.4	5.0*	2.4	5.2	3.2	3.4	3.1
22	4.0	3.4	4.7	2.0	2.4	4.0*	5.0*	4.8*	4.1	4.6*
23	4.5	3.2	8.5*	2.2	2.4	3.0	3.7*	3.5	5.0	3.8
24	3.7	8.0	9.2*	2.0	2.2	8.2*	7.0*	17.0*	21.0*	4.2
Total time	120	77	107	53	68	78	108	91	120	80
Mark obtained	21	18	22	16	18	18	14	18	22	18

3. Problem (23) takes 8.5 seconds giving "bad temper—teacher—blameless." Three years ago *C* was a pupil teacher in a school. The class teacher was bad tempered, often striking children who were blameless. This affected him a good deal.

4. The last problem takes 0.2 seconds. Associations are "Nurse—sister—mother." Last summer his sister cut her foot in the street on some glass. A nurse living nearby had tended it. Last month the nurse fell ill and *C*'s mother repaid the kindness by tending her for three weeks. *C* thinks the nurse fully deserves this courtesy.

TABLE IV

Problems	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	Totals		
											1	2	
1													
2													
3								1	1		2		1
4							1				1		1
5													
6									1		1		
7													
8	1					1		1	1		4		
9					2					2		2	1
10			1		1						2		
11		2	2		2	2		2	1		1	5	1
12													
13	2			1		1				2	2	2	3
14			2			1	1	1			3	1	
15													2
16	2		2			2	2	2	2			6	6
17													
18							1				1		2
19	1	1	1	1		1	1			1	7		6
20	1		2			1	1				3	1	
21	1	1	2		1		2				3	2	1
22						1	1	1		1	4		1
23			1				1				2		
24		2	1			1	1	1	1		5	1	
Total 1	4	2	4	2	3	7	8	5	5	2			
Total 2 .	2	2	5		1	2	2	2	1	2			

There were four type 2 questions. In (14) *C* wondered whether or not thumbs were included as fingers and stopped some time. In (16) he took twelve seconds. He did not know the answer but guessed rightly. In (20) and (21) he had to pause and consider before deciding whether they were true or false.

These two fully worked examples are illustrative of all Very profound associations were met with in some of the tests. *I*, whose average is about four, takes twenty-one seconds over the last question. His mind, he said, was a perfect blank. Analysis revealed a strong sexual complex No person *completely* escaped the effect.

DISCUSSION OF RESULTS

Tables II and IV yield valuable conclusions. Consider Table II. The totals for each individual are at the foot and for each question at the side It is obvious that problems 7, 11, 13, and 16 have a very strong tendency to produce useless association. Eight out of ten persons have type 1 association in question 7. The mean time is 8.8 seconds, while for the next question it is only 3.9 seconds Reference to the list of questions shows these problems are almost certain to bring emotionally toned responses to anyone. Their content is such as probably to affect the average person; therefore this type of question, ought not to be included in the test, since it results in errors This method then acts as a selective agent in improving tests. There are altogether thirty-one of type 1 and twelve of type 2 associations Fourteen mistakes were made (neglecting problem 9) and of these nine are associated with type 1 and four with type 2 delays It was often found that a strong disturbance resulted in a wrong answer as well as delayed response There is very little tendency to perseveration. This may be connected with the fact that a break occurs between each problem. Problem 9 is an amusing reflection on the authors of the test Seven persons gave wrong answers; they insisted that the answer given in the book was wrong; the times were not abnormal. In this test a certain amount of latitude has been allowed for different rates of reading. *G* only gets eight marks, but he is a very slow reader, only finishing ten.

Now consider Table IV. Problem (19) is obviously unsuitable since seven people show associations of type 1 and there are six wrong replies. Question (24) is unsuitable and (8) and (22) are doubtful (11) is of interest since there are five associations of the type 2 and as seen the question is actually somewhat more difficult than the rest for townspeople (but not for country people). Problem (16) has six of the type 2 associations and six wrong answers This is due to absence of knowledge since the examinees are English and the question refers to America. In this test there are six and in Table II there are eight questions which give no abnormalities of any description, *i e*, every one does them fairly quickly and gets them right. These are apparently simple questions and the response time is small for each A sprinkling of type 2 is desirable, helping to eliminate the less capable Nineteen wrong answers were given (if problem (16), which referred to America, is neglected). Of these eleven are type 1 associations and one type 2

CONCLUSION

It is apparent from a general survey of the tables that the association factor is of real significance in affecting the rate of reply in a test In all four hundred questions were set Of these one-hundred four take long times Seventy-one of these are due to type 1 and thirty-three to type 2 associations. There is a marked tendency for wrong replies to occur with type 1 delays Thus thirty-three wrong replies are made, twenty occurring with type 1 delays, and five with type 2. In many cases there is hesitation at the beginning of the test, the first question taking a long time. This is natural, and was not taken as a case of delay. Six out of the ten persons would have answered more questions—and probably earned more marks—had it not been for useless associations Anderson[4] has shown that speed of response in word association diminishes very markedly as the age decreases from fourteen years to eight years. It is thus quite probable that the delay effects will become worse with tests applied to younger children. These delays are present in all mental testing and must affect scores and consequently correlations

The method employed can be made use of as a weapon in the improvement of a test, by selection While the test used is not one of the later improved types, the conclusions are in no way invalidated. For any test which contains similar ideas to these (and most do) will produce associations The results indicate that an attempt should be made to draw up tests, in which a tendency to associate will be improbable To show the effectiveness of this method of examining tests a curve is drawn in figure 1 with the results of *I* and *J* superposed (Table IV). *J* has made no errors due to type 1 and has twenty-two marks, *J*'s time is eighty and *I*'s time one hundred twenty seconds The curves are fairly parallel if the type 1 delays are not counted Thus a drop from (4) to (5), a rise from (15) to (16), then a drop to (17), following with a rise to (18) is common to both curves It is obvious that *I* and *J* are of about the same intelligence and actually if not for type 1 errors both would get twenty-two marks. Questions (8), (10) and (24) are of especial interest, showing how one person takes three and four seconds in a normal response, the other taking ten and twenty seconds in an abnormal response. Emotionally toned

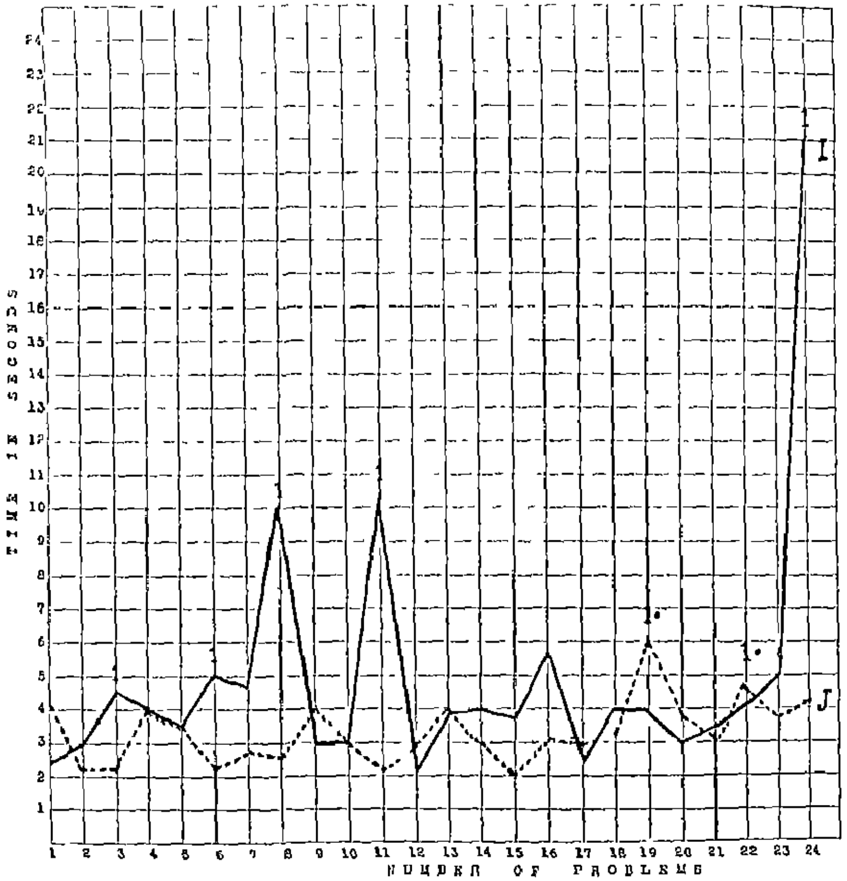


FIG. 1

association, then, can act in two ways: (1) It can reduce the number of questions done, (2) it can produce wrong answers by disturbing the emotions.

To test the effect on correlation, a group of experiments is now being conducted in conjunction with Mr. Vernon Brown. Tests are being given to a hundred children. A group of ordinary tests, involving verbalism and likely to give associations is being given and their correlation determined. To the same children a group of the "illiterate" type test, largely geometrical in nature, is also being given, and their correlation determined Since association is less likely here, the correlation is expected to be higher Results will be communicated later.

References

1 Holzinger *Journal of Educational Psychology*, Vol XIV, 1923, p. 278
2, Stenquist *Journal of Educational Psychology*, Vol. XIII, 1922, p 54.
3 Jung. "Studies in Word Association." 1918.
4 Anderson· *Journal of Educational Psychology*, Vol VIII, 1917, p 97.

TETRAD-DIFFERENCES FOR VERBAL SUBTESTS RELATIVE TO NON-VERBAL SUBTESTS

WILLIAM STEPHENSON

University College, London University

INTRODUCTION

The present paper continues consideration of data gathered from a population of 1037 girls to whom had been applied a "non-verbal" and a "verbal" group test.[1,5]

The Spearman Theory of Two Factors has been shown to fit the non-verbal subtest intercorrelations with exactness,[1] while the verbal subtests only approximately fit the Theory.[5] We observed excess error in the tetrads for the verbal subtests, of amount about 0.015, and the many sources of error brought forward by us were inadequate singly to explain it. At the conclusion of our previous paper,[5] it was hinted that a summation of many small disturbances might be taken to explain the observed 0.015 excess error, but we could suspect, with at least equal reason, that further influences might be at work in the verbal subtests. This latter suspicion, indeed, has some support from data for correlation tables (not reported here) for 100, 200, and 500 sub-populations of the 1037 girls, where the non-verbal subtests showed nearly exact diminution of tetrad error as the sub-population was increased while the verbal subtests uniformly showed excess error in the tetrads, no matter what the size of the sub-population.

However, the present paper is to be concerned with intercorrelation tables containing both non-verbal and verbal subtests, and new facts might be expected to emerge from the comparison that we are to make.

INTERCORRELATIONS AND TETRAD-DIFFERENCES

Table I gives intercorrelations for the seven verbal subtests Nos 2 to 8[5] and the four non-verbal subtests Nos I, III, V, and VIII.[4] All were calculated for standard "normal" score distributions,[1] were calculated by the difference method of scores; and were checked in the usual way, and at the instigation of the tetrad criterion. We have used only four non-verbal subtests instead of the available eight[5] because a table containing all would be unwieldy for tetrad purposes. But in a later table we use other non-verbal subtests.

We note in the first place that the battery of seven verbal subtests correlates 0.65 with the battery of four non-verbal. Assuming that other subtests would provide correlations similar to those in Table I, a battery of very many verbal subtests would correlate 0.82 with a battery of very many non-verbal subtests. There is shown, then, a large measure of concomitance for the verbal and non-verbal abilities, a fact that should not be overlooked when we turn to our main detailed interest in the placing of both types of subtests in tetrads.

We turn at once to this detailed examination of tetrads for Table I. The correlations with age show no discrimination for the verbal compared with the non-verbal subtests (see Table III[4] and II[6]), and we continue without partialling out age correlation. Correction for age correlation usually accounts for at most 0.005 excess error in our tetrads. We note that specificality might be expected for r_{V6}, since subtests V and 6 are of analogy type.

TABLE I.—PRODUCT-MOMENT CORRELATIONS, $N = 1037$ GIRLS

	I	III	V	VIII	2	3	4	5	6	7	8
I		4124	3733	3207	3737	4217	3273	3878	3699	4064	3470
III			4711	3508	3506	3279	3683	3359	3821	3976	4056
V				3278	4602	4145	3908	3826	5029	4985	4797
VIII					2514	2666	2532	2759	3109	3067	3014
2						6013	5580	5085	5676	5845	5883
3							4623	5924	5266	5599	5106
4								4474	4593	5019	5046
5									5550	5745	5121
6										5697	5220
7											5929
8											

There are 990 tetrad-differences for Table I, with value as follows.

Mean of 990 tetrad-differences	0.0355
Observed pe (conventional, mean × 0.845)	0.0300
Theoretical PE (formula 16A[3])	0.0105

Thus, the table shows error of amount 0.028 in excess of that attributable to sampling, and this is not noticeably diminished if tetrads involving r_{V6} are eliminated.

Of the various tetrads entailed, those for only non-verbal subtests give sampling error values, and the tetrads for only verbal subtests entail error of amount 0.018 (in excess of sampling error, as found in the previous paper [5] This latter, spread over the 990 tetrad-differences, would amount to just about 0.002 of the 0.028 excess that is actually observed. The remaining tetrads, to be our concern in the following pages, may be considered for convenience in groups of the following types:

$$r_{n_1v_1}\, r_{n_2n_3} - r_{v_1n_2}\, r_{n_1n_3} \approx f \qquad (z_1)$$
$$r_{nv_1} \cdot r_{v_3v_4} - r_{nv_2} \cdot r_{v_1v_4} \approx f' \qquad (z_2)$$
$$r_{n_1v_1}\, r_{n_2v_2} - r_{n_1v_2}\, r_{n_2v_1} \approx F \qquad (y)$$
$$r_{n_1n_2}\, r_{v_1v_2} - r_{n_1v_1}\, r_{n_2v_2} \approx F' \qquad (x)$$

Where *n* stands for a non-verbal, and *v* for a verbal subtest, the subscripts denoting different subtests, taken for all the combinations possible

Tetrads of Type z_1—Specificality in the "cross" correlations (r_{vn}) will be shown in these tetrads, and it is important to consider this type before the *x*-type, because the "cross" correlations enter critically into *x*-type tetrads The 84 tetrads of type z_1 have the following value·

Mean	0.0241
Observed sigma (× 0.6745)	0.0194
Or, observed pe (mean × 0.846)	0.0204
Theoretical PE approximately	0.0105

We have mentioned that r_{V6} may entail specificality, and the elimination of tetrads involving r_{V6} leaves 76 tetrads, with mean 0.0226, *i.e.*, observed pe 0.0191 (conventional). The z_1 tetrads thus entail error 0.016 in excess of sampling error.

To explain this excess we may try out the various possible disturbers of tetrads considered in previous papers That is, can age correlations, "speed preference," "similarity of relations," group testing effects, calculation mistakes and the like, account for the residual, excess, error, either separately or as a whole?

Eliminating all z_1 tetrads that might be disturbed by "speed" and "similarity of relations," leaves 24 tetrads with mean 0.0173, that is, observed pe 0.0140, the theoretical value being approximately 0.0105, and partialling out the C-measure (the effect of school and class[5]) this mean is reduced to 0.0157, that is, an observed pe 0.0133 The excess error is now only 0.008, and it might be accommodated by age, and calculational mistakes, using the proper sigma, instead of the conventional value (mean × 0.8453), also slightly reduces this 0.008

value. Thus, elimination of one or two possible disturbances would result in the z_1 tetrads showing sampling error only.

Tetrads of Type z_2.—There are 420 tetrad-differences of this type, with value as follows:

Mean of 420 tetrad-differences	0.0253
Observed pe (mean × 0.8453)	0.0214
Theoretical PE approximately .	0.0105

There is excess error of amount about 0.018 in the above tetrads, a value similar to that observed for the verbal subtests.[5] It is obvious that any specificality present in the verbal subtests will be potent also in these z_2 tetrads.

To explain this excess error we may try out all the possible disturbers of tetrads known to us: But the result is the same as that obtained for the verbal subtest intercorrelations, and we are left with excess error of amount about 0.015 (see [5]) that can not receive an explanation except by way of assuming a summation of many slight errors of the kind put forward for the z_1 tetrads.

Tetrads of Type y.—The 126 tetrad-differences of this type have value:

Mean of 126 tetrad-differences . .	0.0171
Observed pe (mean × 0.845) . .	0.0144
Theoretical PE approximately .	0.0105

Omitting the tetrads which involve r_{v6} brings this mean down to 0.0164, *i.e.*, observed pe 0.0138. The excess error, like that for the z_1 tetrads, can be accommodated by age, C-measure, and "speed" effect. Even before making allowance for these two or three influences, the excess error is only 0.009. The y-type tetrads may thus be taken to fit the tetrad theory.

At the present juncture we should ask whether an explanation of the excess error observed in the various tetrads, in terms of a summation of a few slight errors due to effects such as age, calculation mistakes, etc., is reasonable and likely to have a factual basis. On the one hand the non-verbal subtest intercorrelations show no excess error in their tetrads.[1] But the various suggested disturbers of tetrads mentioned above (and in previous papers) are authenticated in other work. The influence of disturbers such as "speed" (for subtests 3, 6, 7), "propinquity" (especially for r_{78}), and idiosyncrasies (for subtests 2 and 5, relative to all others), can be observed by using the verbal subtests 4, 5, 6, and 8, together with the non-verbal subtests: It is then found that (on eliminating tetrads involving r_{v6}) a noticeable diminution is observed in tetrads which, apparently, are not disturbed by influences of "speed," "propinquity," and "idiosyncrasies." But it would require 9 errors, each of magnitude 0.005 (the amount usually found for the influence of age correlation), to cover an observed excess error of 0.015 ($\sqrt{9 \times 0.005^2} = 0.015$). Assuming an excess observed error of 0.005 per influence, for each of age, school and class, "speed," "propinquity," "idiosyncrasies," group testing, "similar relations," and calculation mistakes, a total excess error of amount 0.015 would result. The assumption, however, is perhaps at most only on the border-line of acceptability. Thus it is possible that the 0.015 excess error observed in the case of the verbal subtest intercorrelations,[5] and for the z_2 type tetrads, can not be reasonably accepted as due to a summation of numerous small errors of the kind brought forward, while, on the whole, the excess error shown by the z_1 and y tetrads (less than 0.010 in each case) is possibly reasonably acceptable.

Tetrads of Type x.—This is the last type of tetrads to be examined. The total 252 x-type tetrad-differences have the following value:

Mean of 252 tetrad-differences	0.0703
Observed pe (mean × 0.845) ..	0.0595
Theoretical PE approximately	0.0105

Of the 990 tetrad-differences for Table I, (*i*) all the largest are of x-type, and (*ii*) all but three of the x-type are of positive sign when the tetrads are in the form given at (x) above. (Of course, unless, otherwise stated, no regard is taken of the sign of tetrad-differences, because for each positive difference there is an equal negative difference.)

If we omit tetrads involving r_{16}, 234 x-type tetrads remain, with mean 0.0726. No matter what effects we try to make allowances for we always encounter excess error of amount about 0.059. If we allow 0.015 excess error, attributable to a sum of disturbances of the kind accepted previously for other tetrads, there still remains excess error of amount 0.057 in the x-type tetrads for which we have as yet no explanation.

The contributions of the r_{vv} correlations to the x-tetrads are given in summary form by tetrads of the following type, where r_{vv} is the correlation concerned, taken with all the available non-verbal subtests, two at a time:

$$r_{vv} \cdot r_{n_1 n_2} - r_{vn_1} \cdot r_{vn_2} = f \qquad (w)$$

There are 12 tetrads of this type for each r_{vv}, and *means* of these sets of 12 are given in Table II. The mean for the 12 w-tetrads for r_{23} is 0.0992, for r_{24} is 0.0910, etc. No single correlation among the verbal subtests gives a small difference in these w-tetrads. The excess error seems to cover all the verbal subtest correlations relative to the non-verbal.

Table II.—Means for 12 Tetrad-differences (w-type), for Table I

	2	3	4	5	6	7	8
2		0992	0910	1020	0730	0778	0848
3			0540	1001	0623	0691	0574
4				0548	0427	0558	0020
5					0741	0786	0617
6						0591	0475
7							0699
8							

Consideration of the x-type Error

It is obvious that neither age, calculation mistakes, nor propinquity effects can account for the residual error shown by the x-tetrads Any general effects due to author's idiosyncrasies can not be taken to be potent, because x-tetrads for r_{23}, r_{24}, r_{26}, r_{27} and r_{28}, and the corresponding correlations for subtest 5, should be free from residual error if idiosyncrasies were effective, since subtests 2 and 5 are Thorndike subtests, while the rest are of the author's construction.

Group testing anomalies can not be taken readily to explain the residual error The effect is not noticed in the tetrads of any other type of tetrads It is true, though, that any effect characteristic of verbal subtests *only* (or non-verbal subtests only) would be observable under the critical conditions presented by the x-tetrads The x-type error, however, has been observed previously under conditions free from group testing anomalies,[3] and, further, a test of the influence, made by calculating separate r's for each of eleven testing groups, gave results paralleled by the C-measure correlations (given later in this article), *viz*, that the x-type error remained when the influence was taken into account.

If "speed preference" enters the various subtests it does so in no obvious way. On introspective grounds, from the consideration of errors made in the test-units, and from the general experience with subtests, we must submit that subtests 2, 4, 5, 6, and 8, afford a battery giving good power-speed measure, and that the tetrads involving these subtests are free from gross "speed preference" so far as these subtests are concerned. Of the non-verbal subtests, I, III, and IV, are likewise good power-speed tests, calling for the best that a child can give under conditions free from hurried, slighted, work. The x-tetrads for these subtests show residual error that cannot be differentiated from that for any of the other subtests. If a misbalance of the quantity-quality function is critical in our subtests, then it would have been found in previous work and, certainly, the whole question of the function would require re-experimentation from the foundations upward.

We are left, now, to consider the effect of "school and class" which, we say, may resemble the effects that might be attributed to group testing anomalies.

The Effect of C-measure.—The C-measure[5] is a score given to each girl, the same for each girl in a particular class, on account of school "standing" and class position. It is objective to the extent that its foundation is the order Standards IV to VIIb. It should serve, to some extent, as a measure of scholastic influences, such as reading ability.

The correlations of C-measure with verbal subtests have been given previously[5]. The following new correlations are required $r_{c,\text{II}}$, $r_{c,\text{III}}$, $r_{c,\text{V}}$, $r_{c,\text{VIII}}$ having values 0.3912, 0.3075, 0.3647, 0.2858, respectively. With the correlations now available we can calculate partial correlations for Table I, for C-measure partialled out. The resulting partial intercorrelations show that the non-verbal subtests have average intercorrelation 0.2973, the "cross" correlations have average 0.2700, and the verbal subtests have average intercorrelation 0.4503. It is obvious that the tetrads for such a table will agree, type for type, with those for Table I. The results for the z_1, z_2, and y tetrads, for the n-tetrads (non-verbal subtests among themselves) and v-tetrads (verbal subtests among themselves), are similar to those given above for Table I, except that slightly less residual error is observed throughout. The 252 x-type tetrads have value as follows

Mean, 252 x-type tetrads, Table I, with C-measure partialled out	0.0640
Observed pe (mean × 0.845) . .	0.0541
Theoretical PE approximately . .	0.0105

Omitting tetrads involving r_{v6} (analogy subtests) leaves 234 tetrad-differences, with mean 0.0660: Omitting those for r_{35} and r_{56} in addition, leaves 213 tetrads, mean 0.0644.

If we now allow, for these C-measure partialled correlations of Table I, the 0.015 excess error which might be attributed to a sum of errors due to "speed," "propinquity" (especially for r_{78}), and "idiosyncrasics," age, calculation mistakes, and the like, we have still error of about 0.0512 for the x-type tetrads, in excess of sampling error. The grossest of these disturbances (due to "speed preference" in subtests 3, 6, and 7, to "propinquity" for r_{78}, and to "idiosyncrasics," for 2 and 5, relative to the other verbal subtests) should be absent from x-type tetrads for the four verbal subtests 4, 5, 6, and 8 (if r_{56} is duly accounted for). The x-type tetrads for these four subtests with the four non-verbal subtests, for C-measure partialled out, omitting the tetrads involving r_{56}, are 54 in number, with mean 0.0511—roughly the same as that above for the C-measure partialled table as a whole, allowing 0.015 excess error.

Thus, with what may be described as the refinements entailed by the C-measure partials, and allowing for excess error 0.015 as the sum of that due to various specified disturbances, the x-type tetrads nevertheless show error 0.050 in excess of that attributable to sampling and the specified influences.

A SECOND INTERCORRELATION TABLE

We have used in Table I only part of the correlational data available, and it should be determined whether other intercorrelation tables give results in agreement with those already found in Table I. In Table III we have included six non-verbal subtests and four verbal subtests. The "cross" correlations for the subtests II and IV alone are new.

There are 630 tetrad-differences for Table IV, with the following value.

Mean of 630 tetrad-differences	0.0294
Observed pe (mean × 0.845) ..	0.0248
Theoretical PE approximately .	0.0100

There is observed, once more, appreciable error in excess of that attributable to sampling. We now consider the tetrads in the z_1, z_2, y, and x types.

TABLE III.—PRODUCT-MOMENT CORRELATIONS, N = 1037 GIRLS

	3	4	6	8	I	II	III	IV	V	VIII
3		4023	5200	5100	4217	3001	3270	3807	4145	2066
4			4593	5040	3273	3058	3083	2029	3008	2532
6				5220	3000	3281	3821	3715	5020	3109
8					3470	3192	4050	3753	4707	3014
I						3403	4124	3274	3733	3207
II							3484	2074	3424	3707
III								4140	4711	3508
IV									3771	3012
V										3278
VIII										

There are 240 tetrads of type z_1 for Table III, with mean of amount 0.0222; allowing for r_{V6} and $r_{II,VIII}$ (the former because both V and 6 are Analogy subtests, the latter because of the specificality shown in the first paper[4]) leaves 100 tetrads, with observed pe (conventional, *i.e.*, mean × 0.8453) of amount 0.0136. Excess error of amount 0.0095 is therefore entailed.

There are 72 tetrads of type z_2, with mean 0.0182. Eliminating r_{V6} leaves 66 tetrads, with observed pe of value 0.0142, *i.e.*, an excess error of 0.0100

Some of the y-type tetrads for Table III have been included already in the results for Table I, but the 54 new tetrads of this type have mean 0.0144, or an observed pe of amount 0.0122, for the theoretical PE of amount 0.0100. The excess error is now only 0.007.

Thus, excess error of about 0.0100 is found for the z_1, z_2, and y tetrads. It is obvious that age correlations will reduce this excess, and a sum of slight errors attributable to "school and class," calculational mistakes, and the like, might possibly explain the excess error shown by these tetrads. So far the data confirms that found for Table I.

There are 108 new x-type tetrads for Table III, with mean 0.0520. Eliminating tetrads involving r_{V6} and $r_{II,VIII}$ leaves 90 tetrads, with mean 0.0480. These x-type tetrads thus show excess observed error of amount 0.0340. It is found that approximately this amount of error is associated with each r_{vv} for Table III The w-type tetrads for Table III, for r_{34}, r_{36}, r_{38}, r_{46}, r_{48}, r_{68}, respectively, have mean 0.0505, 0.0554, 0.0516, 0.0448, 0.0615, and 0.0484. Thus no single r_{vv} is associated with small x-type tetrad-differences. The results are again similar to those obtained for Table I. The most significant tetrads are those of x-type. But when subtests II and IV are included in the table of correlations, as in Table III, the various errors concerned are less than those met with for Table I, a fact that should receive consideration for a few moments.

The subtests II and IV correlate with C-measure more highly than the other non-verbal subtests, and are like the verbal subtests in that respect. This may be taken to account for some of the decrease in excess error shown by the 90 x-type tetrads, when compared with the 213 x-type tetrads for Table I. It is a matter of some interest to pay regard to subtests III and V, comparing them with subtests II and IV. On the one hand, it may be suggested, the subtests II and IV are not so "novel" as subtests III and V (or as I and VIII as well); they are not so likely to be so dependent upon fore-practice, not-clear-understanding of the test requirements, conative inhibitions, and the like. It might be considered that effects of this kind have entered significantly into the x-type tetrads for Table I, so explaining the difference in amount of excess error observed for these tetrads in Tables I and III. On the other hand, subtests II and IV can not be said to be of much intrinsic worth as instruments in which eduction is critical: We would place more value on subtests III and V, which are perhaps the best of the non-verbal in point of eduction mechanism. We took some pains, in our first paper,[4] to show that $r_{III,V}$ could not be taken to entail specificality relative to the rest of the non-verbal subtests, and the excess observed error for the x-type tetrads for $r_{III,V}$ is approximately 0.090. Thus, against a possible influence such as "novelty" (including therein influences like not-clear-understanding of the test directions or requirements etc.) we can but place the knowledge that subtests III and V involve eductive mechanism to an extent that can not be vouchsafed for subtests such as II and IV. We are not disposed to accept the view that "novelty" and the like can be taken to account for the greater excess error observed in the x-type tetrads for Table I, chiefly because influences that might be anticipated to be as effectual as "novelty" (such as, say, age effect) do not show excess error so large as this under consideration, namely an amount 0.036, given by $\sqrt{0.050^2 - 0.034^2}$, where 0.050 is the observed excess error in the x-type tetrads for Table I, and 0.034 is that for Table III. However, there is room for further work in which the matter of "novelty," fore-practice, etc., can receive critical attention. The matter, indeed, is to receive attention in some work that we have begun with a 500 population.

Finally, to continue the consideration of Table III, we must conclude that the x-type tetrads show excess error that can not receive adequate explanation in terms of the many influences brought forward by us in the course of this and previous papers The data for Table III is similar to that obtained for Table I

Thus, after taking account of many sources of error known to us, including scale anomalies, the effects of school and class, calculation mistakes, age effects, "speed preference," group testing anomalies, etc, we are left with a not inconsiderable observed excess error in x-type tetrads For all the x-type tetrads that can be constructed for our verbal and non-verbal subtests this excess error is not less than 0.050 in amount, after making allowance for sampling and for an amount 0.015 attributable to influences such as age, etc.

THE PROBLEM OF A VERBAL GROUP FACTOR

A group factor may be shown for the verbal subtest intercorrelations when G_n (the common factor observed in the non-verbal subtests[1]) is partialled out. Using the non-verbal subtests as "reference values," in terms of which the g-saturations of the verbal subtests may be determined,[2] we calculate the *specific correlations*[2] for the verbal subtests among themselves: The results are given in Table IV. The Table is for G_n partialled out. We have to decide whether these specific correlations fit the tetrad criterion, so that a group factor may be taken to run through them.

TABLE IV—PARTIAL CORRELATIONS, FOR G_n PARTIALLED OUT, DERIVED FROM TABLE I

	$2g$	$3g$	$4g$	$5g$	$6g$	$7g$	$8g$
$2g$		418	356	404	316	335	356
$3g$			215	306	252	206	236
$4g$				204	180	233	257
$5g$					311	333	252
$6g$						268	211
$7g$							318
$8g$							

There are 105 tetrad-differences for Table IV, with the following value:

Mean of 105 tetrad-differences	0.0195
Observed pe (mean × 0.845) .	0.0165
Theoretical PE approximately .	0.0100

If we eliminate tetrads which involve r_{36} and r_{66} (which, we have reason to suspect, involve specificality due to similarity of relations), the above mean is reduced to 0.0166, *i.e.*, there is observed pe of amount 0.0139, for a theoretical PE 0.0100. It is obvious that by partialling out age correlation and C-measure this observed pe will become nearly exactly the same as the theoretical value. We conclude that the Two Factor Theory fits the specific correlations. The new factor, which appears to be common to our verbal subtests, may be named *V*.

We have arrived at the conclusion, then, that the verbal subtests involve two common factors, G_n, which we would take to be the universal *g*-factor, and *V*. This may explain the apparently intractable residuum of observed error in the tetrads for the verbal subtest intercorrelations (comparing the results found in our first paper[4] with those for the second paper,[5] for the non-verbal and verbal subtests respectively), and in the z_2 tetrads for, as was suggested earlier in this paper,[*] thee excss error there seemed to be on the border-line of acceptability in terms of extraneous influences of the kind "school and class," age, etc. Now the tetrad effect of *two* common factors becomes scarcely distinguishable from that of only *one* common factor when the correlations are of nearly equal magnitude, and it will be noted that the verbal subtest intercorrelations are of such a nearly equal magnitude. Thus, if we accept the G_n factor found for the non-verbal subtests, and the *V* factor found in addition for the verbal subtest intercorrelations, then the results obtained for the tetrads of the verbal subtest intercorrelations fit well with the theoretical expectations.

We have now to turn to consider explanations of the *V* common factor, of the observed excess error in the z-type tetrads.

If the various effects and conclusions considered in the course of our papers are acceptable, we have inductively narrowed the field of possible explanations for the *V*-factor. We are left, indeed, with a possible explanation in terms of "verbality" It is true, however, that doubts remain, especially concerning the influence of "speed," "novelty" and fore-practice, and conative differences in the verbal and non-verbal subtests.* However, our data are not put forward as a finished product, but as primarily an exercise in correlations, tetrads, and errors other than sampling The verbal and non-verbal subtests used in our work, while being similar to those generally used in "intelligence" tests, are not completely suited to work that would make contact with matters of scientific psychology: We can now employ tests which entail eductive principles more thoroughly, we can construct better non-verbal subtests on primarily perceptual lines. We have under way a re-experimentation with a 500 population, which, it is hoped, will go far to place us correctly on the path of facts of some psychological value It is with knowledge of the limitations, and the need for further work, that we turn to consider our *V*-factor in terms of "verbality."

---

[*] Under Tetrads of Type *Y*.

We should ask, in the first place, how wide the *V*-factor might be expected to extend. Is it likely that it will extend through *all* verbal subtests whatsoever, so constituting a verbal *general* factor? Or is it likely to be confined to the collection of verbal subtests used in our experiment, so constituting a *limited* factor?

We have observed disturbances in x-type tetrads previously,[3] for other verbal and non-verbal subtests. But it is more important to recall that work by Davey,[1] although obtaining results that correspond apparently closely with ours, yet found that results for two verbal subtests did not correspond with the more usual result, indicating instead the absence of any verbal general factor or factors Thus, Davey took the view that the specificality shown by verbal subtests must be of limited range Further, the specificality observed for some of her verbal subtests was attributed to "similarity of relations"

Consideration of Davey's Data and Conclusion—Davey's work was with verbal subtests, orally applied, and pictorial subtests Now pictorial subtests are at best only secondarily perceptual Words and sentences are used in the test directions of all our subtests, but once the purpose of the subtest is "set," a difference between our verbal and non-verbal subtests is that words are not ostensibly used, even implicitly, in the non-verbal subtests (excepting, say, subtest IV). This condition, however, might not have held in Davey's pictorial subtests. Again, some of Davey's pictorial subtests had words or sentences directly involved (pictures had to be selected to match short paragraphs with words omitted, names of items in pictures had to be written down, etc., these verbal parts being orally applied). Thus, there is a possibility that word-entailment, with possible "reproduction" individual differences,[5] is not subjected to sufficiently critical experiment in the pictorial subtests. Some of our non-verbal subtests are more primarily perceptual than pictorial subtests can be.

---

* One possible influence, so far not mentioned, is that of an effect due to the subtests being applied a day apart Previous work, however, shows that such an influence must be slight, at most 0 005 error resulting in x-type tetrads

We have some evidence that the oral subtests used by Davey entail specificality relative to other verbal subtests,[3] and some of the specificality observed by Davey might be attributable to an influence of the oral presentation. The oral Analogies, Opposites, and Classification subtests entail specificality which Davey attributed to "similarity of relations," and our work offers some support for such a specificality in these subtests. But the *V*-factor in our work is observed over and above this latter specificality, and likewise over and above any factor attributable to presentation of the test material. It seems that the *V*-factor is over and above any specificality found by Davey, a matter that would follow upon acceptance of the arguments put forward in the previous, and following, paragraph.

Again, the oral Inferences and Likelihood subtests, which alone of Davey's subtests did not show excess error in the x-type tetrads, and upon which the conclusion concerning a *limited* factor depends, correlated only 0.35 (approximately). Under the conditions most verbal subtests correlate more highly. The test-units in these two subtests are somewhat lengthy and complex in structure, especially perhaps for eight to ten year olds of Davey's eight to fourteen year group. It is possible that not-clear-understanding of the test requirements, or excessive memorization effect,[3] or lack of ability of keep the whole test-unit before the "mind's eye," have entered critically into these two subtests, resulting in a smaller correlation than might be expected in the circumstances.

Thus, the facts are perhaps not too strongly in support of the conclusion concerning a *limited* verbal factor, and nothing hitherto found in this field can be taken to contradict the facts found in our work.

After eliminating disturbances due to various possible influences, our work indicates with some certainty that specificality runs through our verbal subtests, and amounts to a single factor that we have named *V* The specificality would appear to be the more discernible the more the non-verbal subtests tend towards being primarily perceptual. Further experimentation is required before we can decide that the *V*-factor is very extensive, but our work may be taken to tend to indicate that the *V*-factor would possibly extend through most verbal subtests in current tests of "intelligence." If our consideration of Davey's conclusion has any factual basis, then we may conclude that no work extant contradicts the possibility of their being *V*-specificality, approximately a single *V*-factor, of very wide range

The possible contact of this *V*-specificality with "reproduction" has been touched upon in the previous paper,[5] and we would suggest that the facts of the general law of retentivity of dispositions,[3] introspection, knowledge that we have of individual differences in the categories specified,[6] and the facts of aphasia, all tend to support a view that *V*-specificality might be expected as the consequence of individual differences that we have covered by the term "reproduction" *i.e.*, reproduction of words, phrases, sentences, or ideas The theory, in any case, seems worthy of further experimental attention. It is not improbable that "reproduction" constitutes an influence in the "speed preference" effect in verbal subtests.

For the present, then, we would say that it is possible that when reproductive influences are allowed scope in subtests, then *V*-specificality would tend to show. Similarly, no doubt, vocabulary can augment the *V*-specificality; although our data give no indication that the *V*-specificality is conditioned by an antithesis that was employed in the verbal subtests. (Subtests 3, 6, 7 and 8 were constructed of words with simple concrete meaning, while 2, 4, and 5 contained words of more abstract meaning)

In point of fact we obtain *V*-factor extending through all our verbal subtests, tending to indicate a general *V*-factor in addition to the universal *g*-factor in verbal subtests We have mentioned previously, however, that the conclusion is tentative There is room for improvement in the matter of the facts, and we await with interest the results of our second experiment with other verbal and non-verbal subtests

CONCLUSIONS

We have covered a mass of correlational data in the course of two papers [4,5] and the present one, dealing with subtests applied to a 1037 population of girls No doubt brevity has left marks of ambiguity or of unintelligibility, for we have had to give conclusions frequently without much reference to the relevant data. There are likely to be found errors of calculation or transcription But any reworking of the correlational data would, we submit, give us no cause to depart from the following general conclusions.

1. The non-verbal and verbal subtests have a high correlation, amounting to 0 82 for a summed correlation for many subtests of both kinds The fact stands in opposition to the opinions that have sometimes depicted the two abilities as independent

2 In the case of the non-verbal subtests, the tetrad-differences matched with some exactness the value to be expected from sampling error [4] This held good regardless of the size of the sample, for sub-populations of 100, 200, 500, up to the full 1037.

3. In the case of the verbal subtests[6] the tetrad-differences were appreciably larger than those to be expected from sampling alone After making allowances for evident disturbances there remained excess error of amount approximately 0 015, a value scarcely to be accepted as merely due to disturbances of the kind considered in explanation of it

4 But much higher residual error came from the tetrads involving both verbal and non-verbal subtest correlations This was of magnitude about 0 060, thus indicating with certainty some factor or factors in one or both of the two kinds of subtests

5 On closer examination the evidence was against any group factor in the non-verbal subtests, but was in favor of one group factor extending rather evenly throughout the verbal subtests Such a factor, moreover, would explain the roughness of the fit of the tetrad formula to the verbal subtest intercorrelations, as mentioned at (3) already given

6 On the whole, the indications are that this *V*-factor extends through all verbal abilities, and therefore may be called a general factor *V* (as contrasted with the universal general factor *g* which is found in both verbal and non-verbal subtests alike) On this matter, however, there are required further facts These are to receive consideration in a subsequent experiment

In conclusion, we would offer, once more, our very sincere thanks to Professor Spearman for willing guidance, for many aids to a clearer view of facts considered, and for many kindnesses given throughout our work in his Laboratory It has been a true pleasure to have been a student and Research Assistant under Professor Spearman.

REFERENCES

1 DAVEY, C. M , *British Journal Psychology*, Vol. XVII, 1926, p 27

2 SPEARMAN, C "Abilities of Man Their Nature and Measurement" The Macmillan Co , 1927

3. Stephenson, W Thesis, Library of University of London
4 Stephenson, W: Tetrad-differences for Non-verbal Subtests *Journal of Educational Psychology*
5 Stephenson, W Tetrad-differences for Verbal Subtests *Journal of Educational Psychology*

ORGANISMIC PSYCHOLOGY AND EDUCATIONAL THEORY[1]

KENNETH SELTSAM

University of Minnesota

There has, within the last five years, come into prominence a so-called new school of psychology known in its original setting as "Gestalt-theorie," as "Configurational" in the translation of Titchener,[2] and still more recently in the language of Wheeler, as "Organismic."[3] Very few people are aware of the fact that in point of origin, at least, it is almost contemporaneous with "behaviorism," that if anything, it is older of the two.[4] Whatever may be the opposed contention of systematic authorities, the members of the school consider it to have begun with a study of "apparent-movement" made by Wertheimer in a German laboratory some time in 1912 Its longer survival than that of behaviorism, it is said by some, may be attributed to the fact of its relative slow growth However that may be, the organismic thought is demanding serious consideration from psychologists in general It is forcing the perhaps too long unquestioned orthodox psychology to make a more careful invoice of stock And in that probing process, if we may so name it, educational theory has not been, and cannot be, immune

THE MAIN CONTENTIONS OF ORGANISMIC PSYCHOLOGY

Before considering organismic, or gestalt psychology, as it is related to educational theory, it would perhaps be advantageous to inquire into the school's main contentions Roughly they may be classed under two heads: Those relative to the circumstances involved in any event, and secondly, those pertaining primarily to the individual experiencer. To the first of these, would be given the name situation-as-a-whole. A response, the organismic psychologists say, is never made to an isolated stimulus. The $S \rightarrow R$ description of events which is presupposed by the theories of association, attention, and behaviorism is obviously not true to the facts of the case. A response (to state the law of configuration) is made to a situation-as-a-whole, and if to any particular detail, always to that detail in its relation to the other details. The organismic position maintains also that the psychological situation is not different from the chemical or the physical in its general aspects, and is, therefore, governed by similar natural laws.

---

[1] The Editor announces with regret the death of Kenneth Seltsam on November 30, 1930 This paper was prepared by Mr. Seltsam while at the University of Kansas

[2] Helson, H *American Journal of Psychology*, Vol XXXVI, pp 343, 495, Vol. XXXVII, pp 25, 180.

[3] Wheeler, R H "The Science of Psychology"

[4] Boring, E G "History of Experimental Psychology"

One of the natural laws which is seen to have considerable psychological significance is the Law of Least Energy, whose statement for application to the field of psychology is not essentially different from that for other sciences. If movement of energy is always in the direction of a low potential, one immediately comes to realize that such concepts as "Trial and Error" are at best absurd It is, according to this manner of facing the problem, absolutely impossible for an organism to act without accomplishing something with regard to the point of low stress, which organismically is defined in any learning situation as the "goal." In general, the situation-as-a-whole is seen to be an enormously complex thing. As such, it cannot be represented by the tremendously "un-complex" symbol *S*, as that sign has historically been used.

As one might expect, the organismic psychologists have still more to say of the organism-as-a-whole They rebel violently against the atomistic physiology of the conditioned reflex school and others. They maintain that structural analysis can never act as an end in itself; that to assume that psychology is to reach the ranks of a science through the analyses of great introspectionists, as the Titchenerian formula reads, is an impossible assumption. The organism-as-a-whole is to be approached as an integrated unit through the medium of functional analysis. Experimentation is to be conducted by alteration of the conditions; and when this alteration takes the form of extreme limitation (as in the case of neural study) it is to be recognized that the results are to a certain extent, at least, abstractions from the real phenomena. Just as it is impossible to think of a situation-as-a-whole in terms of *S*, so it is, in the light of the organismic position, impossible to label the reaction of the organism-as-a-whole as *R*, simple response.

In general, the major contentions of the new school may be summarized very crudely in what would seem for them a definition of the science Psychology is a science of "wholes" which deals with the responses of organism-as-wholes to situations-as-wholes Anything short of these, on the one hand is essentially physiology, and on the other, physics

THE CRITICISMS OF GESTALT, OR ORGANISMIC, PSYCHOLOGY

The structuralistic psychologists, in the meantime, have not seen fit to accept these violent criticisms as such. One should not expect them to do so One of the typical rejoinders, and one perhaps less logical than any other, is that the new school is, after all, not new One finds such statements as the following from Boring "Like James, Dewey was a *Gestalt-psychologe* twenty years too soon"[1] Squires in a recent comment demands that above all else in our dealings with the new school of thought, we maintain a "balanced historical sense"[2] In a semi-popular discussion, involving about as much over-statement as might be expected of such, Robinson would have us believe that not only is the principle of configuration not new, but of questionable actual significance[3]

The problem of the age of any line of thought is strikingly paradoxical Age is at once both desirable and undesirable Were is not possible to trace organismic thought back through the history of psychology—through Dewey, McDougall, James, the Mills, Stout, and even back to Empedocles in the third century B C—orthodox psychology might have greater ground still for criticism That it has been a development, an evolutionary outcome, rather than a scientific upstart should, it would seem, be a favorable indication It is, also, quite possible to reconcile the fact of age with the idea of a new contribution to make To assume otherwise is to admit that, since "there is nothing new under the sun," it is futile to attempt originality, and that neither sociological nor more narrowly scientific movements have any essential evolutionary value

---

[1] Boring, E G History of Experimental Psychology *American Journal of Psychology*, Vol XLII, Apr, 1930, p 510

[2] Squires, P. C, Gestalt Psychology and the Gestalt Movement, A Criticism of the Configurationist's Interpretation of "Structuralism" *American Journal of Psychology*, Vol XLII, Jan, 1930, p 134

[3] Robinson, E S A Little German Band. *New Republic*, November 27, 1929, p 782

Since a great deal of the organismic experimentation has been conducted with animals,[1] in a manner somewhat strange to traditional method, it is said by some that the gestalt approach is a return to anthropomorphism. What could be, however, in the genuine sense of the word more so than the typical experiment out of which has grown the concept "trial and error" is indeed difficult to imagine. Any experimentation which sets as its criterion of success a performance demanding of an animal at least average human intelligence is ultra-anthropomorphic. If, however, we are to hold that an experiment which gives the animal an "even chance" is thus to be described, it is possible that after all being anthropomorphic is not the "crime" it once was.

Again, we are told that gestalt, or organismic, psychology is purely theoretical—the mind-child, so to speak, of over-zealous Teutons. Since nothing has been proved (so the criticism runs) very little concern need be had. This again seems hardly fair. Were we to assume such an absolute standard in our dealings with scientific facts and methods, progress would be permanently halted. Experimentation of any kind presupposes a certain number of working hypotheses. However one may use logic, it is difficult to imagine an experimental situation not demanding a certain viewpoint of approach which, of necessity, must remain in the theoretical realm. To be sure, any working hypothesis deserves to stand or fall in terms of experimental verification. Until that time, however, to say that certain assumptions are valueless simply because they have not been tested is at best a misconception.

Still again the gestalt contributions have been said to exist essentially in an alteration of concepts. Instead of "trial and error," "apperception" or "libido" we have "insight." "Maturation" has taken the place of the "stamping-in process." And so forth. It is true that the organismic psychologists might give the name "conditioning" or "learning by repetition" to certain aspects of the learning situation they would describe. So, too, one might call one's office-chair "water." In the sense that there are certain fundamental finite constituents such a statement would be correct It would, nevertheless, *not* be true historically And, after all, a concept has value only insofar as there is not contradiction in its historical usage. Essentially, what lies behind such criticisms as those of Kuo, who would so to speak, "junk" all concepts, is a dissatisfaction with the tendency to so jumble their meanings as to make them practically useless.[1] A concept, strictly speaking, must suggest subtle differences and to do so it can not be so overloaded as to mean everything. As such, it is, prematurely, destined to mean nothing

---

[1] Kohler, W., "Mentality of Apes", Wheeler and Perkins. Configurational Learning in the Goldfish. *Comparative Psychology Monograph*, Vol. VII, No. 1 March, 1930; Helson, H.: Insight in the White Rat. *Journal of Experimental Psychology*, Vol. X, 1927, pp. 378-396, Lewis, M. H. Elemental versus Configurational Response in the Chick. *Journal of Experimental Psychology*, Vol. XIII, No. 1

There have been still other criticisms.[2] Generally speaking, each involves a certain validity It would be foolish to contend that gestalt or organismic psychology is in any sense a perfect creation, against which all criticism is misplaced As a school it *is* both old and new, theoretical, and perhaps over-ambitious So long, however, as it has an amplification and a freshness of approach, no true scientist can well overlook its attempts

ORGANISMIC PSYCHOLOGY AS A CONTRIBUTION TO EDUCATIONAL PHILOSOPHY

The discussion so far has dealt rather summarily with the controversial aspects of "Gestalt-theorie" as a school of general psychology. It shall now be the attempt to analyze what seem to be its contributions to educational theory. Such an analysis must, for the most part, be based upon opinion and logic, since educational experimentation in the field is almost non-existent.

---

[1] Kuo, Z Y *Psychological Review*, Vol. XXIX, p 344

[2] Calkins, M W. Critical Comments on the Gestalt-theorie *Psychological Review*, Vol XXIII, 1926, pp 135–158

Hsiao, H H A Suggestive Review of Gestalt Psychology *Psychological Review*, Vol XXXV, 1928, pp 136–141

Pillsbury, W B. Gestalt versus Concept as a Principle of Explanation in Psychology *Journal of Abnormal and Social Psychology*, Vol XXI, 1926, pp 14–18

Rignano, E The Psychological Theory of Form *Psychological Review*, Vol VIII, 1925, pp 118–135

Lund, F H The Phantasy of Gestalt *Journal of General Psychology*, July, 1929

Woodworth suggests that the $S \rightarrow R$ description be changed to read $S \rightarrow \text{org} \rightarrow R$ The organismists of course maintain that this description is still extremely atomistic. It is nevertheless, something of a recognition of the organism-as-a-whole.

In education, if any place, it should be desirable to consider stimulating circumstances as wholes We may, perhaps, owe the development of the behavioristic concept "stimulus," in the extreme isolated sense, to such formulation as Morgan's law of parsimony in science Forever, we have been attempting to "strip" experience to the limit. The lay tendency in this direction is well shown by a study of the behavioristic reactions of cattle made some time ago by Stratton[1] In this analysis, it becomes apparent that there is no correlation between the presence of a red flag and anger in cattle The anger reaction is aroused only when the total situation is established—when the flag is waved in an annoying manner, when the waver jumps about, making peculiar vocal sounds and stirring into motion the dust about him

The language field again provides an interesting source for examples of the inadequacy of the "simple stimulus" theory If I present the words "Blanche White Weiss" in a free association test, to a group of subjects including different ages and professions, the reactions are certain to vary tremendously. To the naive, "Blanche White Weiss" is not more than a rather peculiar name To the person well trained in languages, it seems as three repetitions of the same word, "White White White." Between these two extremes and varying directly with knowledge of French, German and English, will be intermediate reactions Yet, throughout, the physical stimulus is exactly the same.

Until the teacher comes to appreciate the fact that the stimulus which he presents exists only in relation to the learning situation-as-a-whole, the reactions he observes will represent to greater or less degree, a dilemma Only when he realizes that a word of sarcasm, a smile of contempt, or an unfair concession is bound in most cases to emerge from that whole and dominate it, will he be aware of the inadequacy of independent emphasis upon some bit of subject-matter His ability to understand why certain "stimuli" fail to call forth certain "responses," will vary directly with his comprehension of a tremendously complex situation He is not, after all, the first violinist in a metropolitan orchestra. He is a whole orchestra in himself, and while one sound, that of his one arm violin, may dominate, it does so only as an emerged part of a situation-as-a-whole

---

[1] Stratton, G M.: The Color Red and the Anger of Cattle *Psychological Review*, Vol XXX, p 321.

However prone we may be to express disapproval of what historically was "faculty" psychology, we have continued in a striking degree to do "faculty thinking." We talk of "reflexes" and of "automatic habits" We continue to debate the questions as to just what *parts* of our organizations are native and what acquired. In our attempt to be scientific, we have seen fit to study anatomy and physics as atomistic sciences, instead of the psychology of learning A child has been educated in history and geometry, not in general understanding His biology teacher has refused to aid in the correction of his English A premium has been placed upon his accumulation of an abundance of rather meaningless data, even by methods admittedly unethical It seems we seldom stop to think that we are educating a whole child, who exists as an intact organism and only as such

We continue to define learning as "little more than the establishing of bonds between responses and stimuli,"[1] apparently naive to the discoveries of such experiments as those of Haller in the eighteenth century,[2] and the masterly studies of Lashley in our own period[3] A stimulus has been considered to effect certain specific connections, not the whole organism With Gates, we have said, "When once a desired reaction is made, it may be stamped in by further exercise until well learned, so that later it occurs promptly and surely"[4] Learning, through experience as such, or drill, has been emphasized "At present the attitude of many teachers toward drill work is undergoing a wholesome change We are coming back to drill. The neglect of necessary drill work has been one of the bases for just criticism of the 'soft pedagogy' in the schools"[5] All the while we have been speaking with confidence of something of which, as Lashley tells us, we know very little

At the same time, the human organism has been made to act in a fashion foreign to any other natural phenomenon. He has been said to learn through a meaningless repetition of muscular movements Vaguely, the idea of goal-seeking, or action toward a point of low organismic tension, has been recognized for a long time For Kilpatrick, it is "mind-set-to-an-end" He also suggests that before reaching that end the organism may, in the process of choosing certain means "be torn within."[1] Almost nowhere, however, is a cue given to an understanding of the organism as a part of a natural scheme, all of which, it is rather logical to suppose, is governed by comparable laws

---

[1] Douglass, H R "Modern Methods in High School Teaching" Pp 22

[2] Haller concluded that by the time a man died, he would, according to the trace physiology, have to have 200,000 impressions for every gram of grey-matter

[3] Lashley, K "Brain Mechanisms and Intelligence"

[4] Gates, A I "Psychology for Students of Education" Pp 212

[5] Stormzand, M J, "Progressive Methods of Teaching" Pp 227.

We have stumbled time after time in our attempt to explain why it is that John Jones, who has practically the same mental capacity (as measured by certain criteria) as Sam Smith, does not see fit to "apply" himself We have said, "Sam likes to study." Never, apparently, have we appreciated thoroughly the fact that possibly there is just as even a balance of tension in one case as in the other For Sam Smith, study provides the greatest resolution of a most complex system of stresses, internal and otherwise John Jones, on the other hand, is organized differently He had different goals, a different configuration of stresses, and consequently is a different whole. Both individuals function according to the same general principle—organismically, the law of least energy. The thought here is not new. In every language there are proverbs expressing in a crude way the idea involved. But as a point of scientific emphasis in psychology, it is relatively new and can, if correctly appreciated, lead to a great clarification in character and personality study

Finally, organismic psychology forces educational philosophy to reconsider whether the hedonic conception, which has abounded since Socrates and no doubt even before, is after all, the true conception Is "happiness" a goal of man's activity, as such? Does "practice with satisfaction" explain selection for continued usage? Do the very possibilities of attaining happiness fade immediately as that most abstract of all abstract things is aimed toward? If education is to answer negatively with the gestalt or organismic psychologists to these questions, the whole view of education will be more or less altered. It will be broadened The organism to be educated will be seen as an emerged part of a larger whole which in turn, stage by stage, may be said to include the whole universe The organismic psychologist's "human" is a larger being. He is not an isolated phenomenon depending for his growth upon some philosophically abstract principle He moves for the same reason that a current of air moves. For him, experience is hearing, seeing, and feeling all at once He does not see without hearing, nor does he hear without seeing He matures in the same way that any physical body changes He is a being "attuned to the world," and understandable only as such.

---

[1] Kilpatrick, W H: "Foundations of Method." Pp 163

In conclusion, it would seem that we can believe with Boring that:

> The progress of thought is gradual, and the enunciation of a "new" crucial principle in science is never more than an event that follows naturally upon its antecedents and leads presently to unforeseen consequents.[1]

and still agree with Squires:

> Every system has had something of value to contribute to our total fund of knowledge [2]

If gestalt or organismic psychology has helped to clarify our picture of the human organism that we would educate, if it has stimulated inquiry and application (and one has but to examine journals in speech education,[3] music,[4] and so forth[5] to realize the fact) certainly we must recognize that the organismic school has a contribution to make to the theory and practice of education which can not well be ignored.

---

[1] Boring, E G The Gestalt Psychology and the Gestalt Movement *American Journal of Psychology*, Vol XLII, No 2, April, 1930, p 308

[2] Squires, P C. A Criticism of the Configurationist's Interpretation of "Structuralism" *American Journal of Psychology*, Vol XLII, No 1, January, 1930

[3] Gray, G W. Gestalt, Behavior and Speech *Quarterly Journal of Speech*, Vol XIV, Nov, 1928, pp 530–534; Gestalt Again Vol XV, Feb, 1929, pp 85–92.

Parrish, W M Implication of Gestalt Psychology *Quarterly Journal of Speech*, Feb, 1929

Woolbert, C H Psychology from the Standpoint of the Speech Teacher *Loc cit*, Feb, 1930

[4] Webber, R A Good Practice Judgment *Violinist*, Vol. XLIX, No 6, June, 1929, p. 190

Kwalwasser, J Music Appreciation, Is It Vital? *Music Supervisor's Journal*, Vol XVI, No 4, Mar, 1930

[5] Moore, R and Van Waters, M The Child and the New Psychology *Libraries*, Vol XXXV, Mar, 1930, pp 117–120

A GROUP INTELLIGENCE TEST SUITABLE FOR YOUNGER DEAF CHILDREN

R. PINTNER

Teachers College, Columbia University

Up to the present time no group intelligence test has been suitable for the deaf child in the beginning grades of deaf schools. All of the group intelligence tests for the beginning grades in hearing schools, such as the Detroit First Grade, the Otis Primary, the Pintner-Cunningham, etc., make use of verbal directions in giving the test, and hence are not suitable for the deaf. On the other hand, the Pintner Non-language Mental Test, which has proved to be well-adapted for deaf children is too difficult for deaf children below the age of nine or ten. Hence the intelligence of young deaf children from about age five to nine has not so far been measurable by means of group tests. The new Pintner Primary Non-language Mental Test[1] would seem to fill this gap.

This primary non-language test was constructed for the measurement of intelligence below the range covered by the Pintner Non-language Test, which is of little or no use below Grade II. Like the latter, it is non-verbal in content and makes no use of language in the directions. These are carried out by means of pantomime and by examples of procedure performed on the blackboard. It was constructed as a non-language test for kindergarten, Grade I and Grade II, so as to counteract any language handicap that may be present among young children coming from homes of varied amounts of familiarity with the English language. A recent try-out of this test in two deaf schools would seem to show that it has decided possibilities as a group test for the young deaf school child.

The test itself consists of four sub-tests. In the first the subject has to note the object or picture held up by the examiner and then find this object among the pictures in the test booklet and mark it. There are nine such items. Sub-test two calls for the completion of a geometrical form, the correct form being always in view as a guide to the subject. The third sub-test requires the completion of a face, each item having various parts of the face omitted. The last sub-test requires the subject to note the position of the arms of the examiner and then draw them in the correct position on the appropriate picture of a manikin in the test booklet The test is difficult to give and requires much preliminary rehearsal on the part of the examiner, but when given properly it can be understood by children entering school for the first time either in the kindergarten or first grade.

---

[1] Published by the Bureau of Publications, Teachers College, Columbia University.

This test was given to the lower grades in two schools for the deaf for the purpose of testing its suitability for the young deaf child.[1] Practically all of the children in the lowest grades in both schools were tested and a sampling of older children in order to see how far up the test would be discriminative In all two hundred ninety-three children were examined, ranging in age from age four to age fifteen. Table I shows the frequency distribution of the scores by age for all children tested. Up to age eight inclusive the cases are fairly well scattered over the total range of scores; but from age nine onwards the cases begin to be bunched at the upper end The median scores rise steadily from age four to age nine, where they reach a level It is obvious, therefore, that the test is too easy for age nine and above The lack of zero scores at any age would seem to indicate that the test is not too difficult for the ordinary deaf child during the first few years in school It would seem, then, that we have here a test suitable for the young deaf child between the ages of four to eight, a period of development not covered at present by any other suitable non-language group intelligence test The content of the test proved of interest to the children The teachers of the deaf children commented upon the suitability of the content and presentation of the test for deaf children

So far this test has not been adequately standardized on hearing children, but a comparison of the deaf children's scores with the tentative norms for the hearing may be of interest The results up to the present time are as follows.

CA	5-6	6-6	7-6	8-6
Median Scores				
Hearing	11	45	55	65
Deaf	31	49	56	61

---

[1] The writer wishes to acknowledge the splendid cooperation of Dr Harris Taylor of the Institute for the Improved Instruction of Deaf Mutes, and of Miss Carrie Kearns of P S 47, Manhattan Two graduate students, Mr Chester Bennett and Miss Elsa Richards had charge of the testing and the writer is indebted to them for their careful work

The high score for the five-year-old deaf is based upon only nine cases and these probably represent a very superior group. At the other three ages, where we have a better sampling of deaf children the scores are about the same as those for hearing children. This result is in marked contrast to most comparisons between the deaf and hearing on group intelligence tests, where the deaf are usually very far behind the hearing. It may be that our tentative norms for the hearing are too low, or that the deaf children tested are a much superior sampling to deaf children in general. Further work with both deaf and hearing children will clear up these points.

Table I.—Distribution of Scores by Age Groups

Age

Scores	4	5	6	7	8	9	10	11	12	13	14	15	
70-6				3	3	9	8	9	8	1	3	2	
65-9			1	1	12	21	20	14	10	5	2	2	
60-4			1	4	13	13	16	5	2	1		1	
55-9			7	8	11	4	3	2		1			
50-4		1	6	8	6	1							
45-9			3	2	4	1	1	1			1		
40-4		2	2	1			1						
35-9		1	2	1									
30-4	1	2	2	2									
25-9	1		1										
20-4			1	1									
15-19		1	1										
10-14	2	1											
5-9	1	1	1										
0-4													
Total	5	9	28	34	49	49	49	31	20	8	6	5	293
Median	13	31	49	56	61	66	65	68	68	66	71	68	

There is another possible explanation for this equality or superiority of the deaf child on this test. One of the examiners, Mr. Bennett, believes that the deaf child has an advantage over the hearing child inasmuch as his whole training teaches him to pay more attention to gestures, facial expression and the like, and hence he is better able to interpret the pantomimic instructions of the examiner and profit more from them than the hearing child does.

SUMMARY

A group intelligence test for young children has been constructed which is presented by means of pantomime and examples on the blackboard. No language is required to understand the directions or t respond to the test items. By means of this test it is possible to tes young deaf children in schools for the deaf It is the first group tes suitable for such cases. The results so far obtained would seem t indicate that it is discriminative for ages four to eight inclusive. Th deaf children so far tested slightly exceed the tentative norms for th hearing

WHAT IS MEANT BY A *G* FACTOR?

TRUMAN L. KELLEY

Harvard University

The purpose of this article is to discuss what constitutes the proof that a common factor is sufficient to account for the intercorrelations between a number of variables. The particular occasion for it is the article by Professor Holzinger[1] in which he uses certain data of mine as follows:

> By way of illustrating these points, we may take an example from Professor Kelley's book (p 97ff) The four tests used are
>
> X_1 = Reading speed
> X_2 = Arithmetic power
> X_3 = Memory for words
> X_4 = Memory for meaningful symbols
>
> The intercorrelation and tetrad differences are as taken from a paper by the writer.[2]
>
> | | X_1 | X_2 | X_3 | X_4 |
> |---|---|---|---|---|
> | X_1 | | 0586 | 1950 | 2089 |
> | X_2 | | | 1487 | 2480 |
> | X_3 | | | | 6693 |
>
> $t_{1234} = -\ 010 \pm\ 037$
> $t_{1243} = -\ 005 \pm\ 037$
> $t_{1342} = \ 005 \pm\ 016$
>
> From the insignificance of these tetrads we may conclude that factor pattern (1) with only *g* common is sufficiently complex. If group factors are present in these four tests their effect is insignificant, yet Professor Kelley employs the pattern given by the following portion of his Table XII (*op cit*). Numbers in the table are standard deviations

---

[1] Holzinger, K. J Thorndike's CAVD Is Full of *G* *The Journal of Educational Psychology*, March, 1931

[2] Holzinger, K. J. On Tetrad Differences of Overlapping Variables *Journal of Educational Psychology*, Feb , 1929

Tests	α = heterogeneity, maturity, sex, race	β = verbal factor	γ = number factor	δ = memory factor	ϵ = spacial factor	ζ = speed factor	Specific	
							Not chance	Chance
1 Reading speed	40	60			00	38	36	28
2 Arithmetic power	21		63		31		16	66
3 Memory for words	06	60		56			30	33
4 Memory for meaningful symbols	50			52	30		32	30

This happens to be a very false presentation of my argument and findings because the other elements in the problem have not been taken into consideration It is totally unsound to investigate four variables, find that a single underlying factor is sufficient to account for the intercorrelations, and conclude that the same underlying factor (or any other, for that matter) would account for the intercorrelations of these variables when employed in connection with variables other than the four Awareness of this is fundamental to a study of mental analysis, and that Holzinger has overlooked it both in connection with Thorndike's interpretation of CAVD and my interpretation of my tests is unfortunate because it confuses the issues involved

A brief presentation of hypothetical data will, I hope, make the issue clear Let us be given the following eight tests and let them be totally accounted for by five underlying independent, *i e.*, uncorrelated factors labelled *a*, *b*, *c*, *d*, *e*, as shown in Table I The constants are designated by literal symbols, other than x, with subscripts The chance factors that ordinarily enter into measures have been omitted merely for the sake of simplicity of presentation

TABLE I

$$
\begin{aligned}
x_1 &= m_1a + n_1b \\
x_2 &= m_2a \qquad\qquad + o_2c \\
x_3 &= m_3a \qquad\qquad\qquad\qquad + p_3d \\
x_4 &= m_4a \qquad\qquad\qquad\qquad\qquad\qquad + q_4e \\
x_5 &= m_5a + n_5b \\
x_6 &= \qquad\;\; + n_6b + o_6c \\
x_7 &= \qquad\;\; + n_7b \qquad\qquad + p_7d \\
x_8 &= \qquad\;\; + n_8b \qquad\qquad\qquad\qquad + q_8e
\end{aligned}
$$

With variables 1 to 8, due to underlying factors as given, it can readily be shown that the tetrads involving variables 1, 2, 3, 4, will equal zero, implying that one underlying factor is sufficient to account for the intercorrelations This is true but the intercorrelations accounted for are r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34}, and none other. The correlation r_{15} is not accounted for Variable 1 has just as much "right" to be considered in connection with variable 5 as in connection with variables 1, 2, 3 and 4 Dr Holzinger's use of my data is equivalent to abstracting from a table, such as Table I, the first four rows only, and then pointing to the lack of evidence of factors *b*, *c*, *d*, and *e* Certainly there is lack of evidence in the first four rows, but there is no lack of evidence in the table entire

Consider the data of Table I All of the tetrads involving x_1, x_2, x_3, x_4 equal zero, and secondly all the tetrads involving x_5, x_6, x_7, x_8, also equal zero. Clearly one underlying factor is sufficient to account for the correlations between the first four variables Let us call this factor g. Clearly one underlying factor is sufficient to account for the correlations between the last four variables. Let us call this factor g. The next thing is to say that $g = g$. As a bit of verbalism it sounds reasonable but is, in truth, absurd. The fact that one investigator finds a single underlying factor sufficient in a given set of tests, and a second finds a single underlying factor sufficient in a different set, is not evidence at all that the two factors are the same In the Table I illustration they are entirely different, being an *a* factor in the first four tests, and a *b* factor in the last four. To generalize from the relationships found in four tests to relationships supposed to lie in the mind of man begs the question, for doing so involves the assumption that the four tests sample the entire mental life. There is no certain technique which reveals the latter, but the more extensive in number and varied in nature the tests employed, the more likelihood that the relationships found to underlie tests scores will reflect relationships in mental life.

THE GROWTH OF SOCIAL PERCEPTION IN DIFFERENT RACIAL GROUPS[1]

W N KELLOGG AND B. M EAGLESON

Indiana University

I Outline of the Problem

Of the numerous attempts to uncover racial differences between negroes and whites, the possibility of a distinction in the ability to interpret facial expression seems particularly fruitful It may be argued for example that negroes should be less able to understand the emotional expression of *white individuals* because they have fewer opportunities as a class to observe whites in emotional situations. They would correspondingly be supposed to be superior, however, in comprehending the expressions of the negro countenance. On the other hand, an ability surpassing that of whites in the interpretation of emotional expression of all sorts might theoretically be ascribed to the negro as an outgrowth of the greater propensities for emotional behavior which he is generally supposed to possess

One of the most direct methods of approaching such a problem is to determine at what stage of development differences, if any, begin to appear. This can be done by examining groups of children of each race, at varying age levels, equated as nearly as possible in respect to intelligence and social status. G. S Gates[2] has already reported an investigation of the growth of social perception in four hundred fifty-eight white children from three to fourteen years old Since her method is simple and the groups are carefully described we have taken her results as representative for white children, and have duplicated the experiment as nearly as possible upon similar groups of negroes. The technique adopted by Gates has been followed meticulously in order to make the findings validly comparable to those she reports The only factor which remains uncontrolled in our procedure is the geographical difference in the locations of the negroes and the whites That this has introduced no spurious influences will appear, we think, from the results.

---

[1] The writers wish to acknowledge the generous cooperation of the following persons, whose assistance in granting permission to enter the various schools and whose suggestions in the selection of different social groups are sincerely appreciated. Mr Daniel T Weir, Acting Superintendent of Indianapolis Public Schools, Mrs Grace L. Brown, Superintendent of Indianapolis Kindergartens, Messrs. W N Grubbs and E. W Diggs, principals respectively of Public Schools 24 and 42 of Indianapolis, and Mr Anthony Courtney, Principal of the Banneker (negro) Public School in Bloomington, Indiana

[2] Gates, G S: An Experimental Study of the Growth of Social Perception. *Journal of Educational Psychology*, Vol XIV, 1923, pp 449–461

II. EXPERIMENTAL CONDITIONS

Materials—Six pictures of the face, head and shoulders of a woman in various emotional poses were selected from the series published by Ruckmick[1] Typical interpretations given by adults to each of these pictures are as follows[2]

Picture A—Laughter, joy or amusement
Picture B—Pain.
Picture C—Anger or defiance
Picture D—Fear or horror
Picture E—Scorn, contempt or disdain
Picture F—Surprise, wonder or amazement

The prints were presented "singly and individually" by a negress experimenter. Race and sex relationships between experimenter and subjects were thus equivalent to those between Gates' subjects and experimenters It should be noted, however, that the emotional pictures were identical in both studies and that they were *poses of a white woman*.

Method—After preliminary efforts to put the subject at ease, his name and age were asked School grade and sex were recorded at the same time. The experimenter then said (1). "I am going to show you some pictures of a lady, and I want you to tell me what she is doing." If no response, or if an incorrect or doubtful response was made the subject was asked (2): "What is she thinking about?" and, after a pause (3): "How does she feel?" Replies to these questions were recorded verbatim.

The problem of grading and classifying the answers accurately, particularly those of the younger children is, as Gates has pointed out, a difficult one. We endeavored, however, to adhere strictly to the criteria established by her Those responses listed as correct or incorrect for white children were similarly classified throughout this experiment A few answers were received, however, which Gates does not report, in these cases "any interpretation which expressed the general trend of feeling in the picture was counted correct "

---

[1] Ruckmick, C. A: A Preliminary Study of the Emotions *Psychological Monographs*, 1921, Vol XXX, No 136 (Critical and Experimental Studies in Psychology from the University of Illinois), pp. 30-35.

[2] *Cf* either Ruckmick or Gates: *Op cit.*

In scoring, the principle of giving the subject the benefit of the doubt was adopted Thus when it had been necessary to ask two or three of the experimental questions the answer to one of which proved acceptable, the subject was allowed credit for his best answer unless replies to the remaining questions clearly indicated that he had no comprehension whatsoever of the expression pictured

Subjects —Three hundred thirty-two negro school children whose ages ranged from three to fourteen years were selected for this study from as widely varying social levels as were available in the vicinity of Indiana University The total number employed was divided about equally between boys and girls Ninety-three of the group, representing all social strata, came from the only negro public school in the city of Bloomington, Indiana, a community of approximately 18,000 population Of the remaining subjects, ninety-eight were taken from Public School No 42 in Indianapolis,[1] which is located in a middle and upper class negro district A third group, numbering one hundred eight, came from Public School No 24 of Indianapolis, which was chosen as representative of the middle and lower strata of colored families. Children of ages three, four and five were tested in the Flanner House Kindergarten and Day Nursery and in Kindergarten No 4, both of Indianapolis A rough classification according to intellectual ability, as estimated by teachers, was obtained for all subjects six years and over in age

By these selections the authors feel that a fair sampling of various social and environmental levels of the typical urban colored population has been obtained

III RESULTS

Comparisons between Racial Groups.—The summarized findings for all subjects, regardless of social or intellectual status are given in

---

[1] All the schools visited were exclusively negro institutions, including teachers and principals

Table I. Figures of this table are the per cents of correct interpretations, picture by picture, arranged according to the ages of the respective subjects

Aside from minor variations the data indicate substantially the same facts brought out in Gates' study of white children, namely,

TABLE I — PER CENTS OF CORRECT RESPONSES PER PICTURE FOR THREE HUNDRED THIRTY-TWO NEGRO CHILDREN
(All Groups Combined)

Ages . .	3	4	5	6	7	8	9	10	11	12	13	14
Picture A (laughter)	.00	75	.80	87	73	83	.07	1 00	88	94	85	91
Picture B (pain)	00	00	45	.50	68	70	80	91	72	.01	.74	83
Picture C (anger)	00	.13	.20	47	.51	00	02	70	78	87	88	.66
Picture D (fear)	00	00	.15	.03	35	37	51	67	72	74	85	74
Picture E (scorn)	00	00	00	.00	03	03	00	00	13	16	.18	.17
Picture F (surprise)	00	00	00	00	00	03	.11	.12	25	35	29	34
Number of cases	5	8	20	30	37	30	37	33	32	31	34	35

(1) that the percentage of successful responses tends to increase with age regardless of the emotional expression judged, and (2) that the general order of perceptibility is, with occasional exceptions, as follows: Laughter, pain, anger, fear, surprise and scorn. The percentages of successful responses of the negro children exceed those of the white children (as shown in Gates' Table V, p 460) more times in the interpretation of the Picture D (fear) than with any of the other expressions. It is doubtful, however, if any particular significance should be ascribed to this result. The findings on the whole are strikingly similar in the two experiments

One point of some significance which appears in the figures of Table I is the tendency of the percentages to drop for many of the pictures at ages thirteen and fourteen. An examination of the analogous data for white children brings out the same sort of decline at the upper ages. The drop is probably to be accounted for, we think, by the fact that the older children who remain in school (whether white or colored) represent the lower intelligence levels for their ages, and would consequently be expected to be inferior in social perception when compared with subjects slightly younger than themselves

In Table II we have recomputed Gates' percentages for white children, combining the results for all six prints, *i e*, without consideration for the emotion depicted. These data are here paired with the similar figures of the negro children.[1]

TABLE II.—PER CENTS OF SUCCESSFUL RESPONSES OF WHITES AND NEGROES (All Pictures Combined)

Ages	3	4	5	6	7	8	9	10	11	12	13	14
Whites (from Gates)												
Per cent	25	26	28	35	39	39	49	58	73	75	60	77
Number of cases	10	40	85	59	55	58	39	28	44	27	17	8
Negroes												
Per cent	00	15	27	31	38	43	51	57	.58	61	63	61
Number of cases	5	8	20	30	37	30	37	33	32	31	34	35

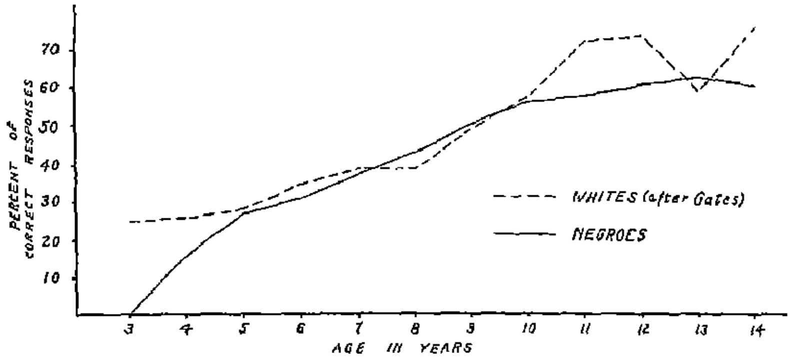


FIG. 1.—Growth curves showing scores of negroes and whites as given in Table II. The scores of the whites are not consistently superior to those of the negroes except at the younger ages where the white subjects were highly selected (see text).

Disregarding for the moment, the percentages below year five, we see that in six of the remaining ten age-groups, the values obtained by the colored children either excel those of the whites or approach to within two points of the results of the whites. There appears therefore to be no obvious tendency for the samples of either race to be superior to those of the other in this ability.[2] The percentages for years three and four are not validly comparable since our figures were obtained from average colored children while the majority of Gates' subjects of these ages came from "the Kindergarten of a select private school."

---

[1] The percentages were computed by dividing the actual successes, regardless of the type of expression judged, by the total possible successes (= the number of cases × 6).

[2] The reliabilities of these differences were not obtained.

Intellectual, Sex, and Social Classifications—Percentages in Table III (*vide infra*) were obtained in the same manner as those of Table II, except that the data were subdivided into three classes according to teachers' estimates of the intelligence of the children

TABLE III—PER CENTS OF SUCCESSES OF GROUPS CLASSIFIED ACCORDING TO INTELLIGENCE

(All Schools Combined)

Ages	6	7	8	9	10	11	12	13	14
Group X (superior)									
Per cent	33	44	46	53	56	65	62	73	67
N	8	12	9	12	13	13	10	10	10
Group Y (average)									
Per cent	31	44	50	54	56	58	61	61	65
N	12	11	10	15	9	11	12	12	10
Group Z (inferior)									
Per cent	30	29	33	45	58	46	61	57	54
N	10	14	11	10	11	8	9	12	15

Group Z, with few exceptions, is lower in social perception at all ages than either Groups X or Y Group X shows little superiority over Group Y except in the later ages, namely, eleven, twelve, thirteen and fourteen. There thus appears to be a positive relationship of a crude sort indicated in our figures between social perception and intelligence. The failure of the X group to demonstrate its superiority over the Y group at the younger ages is probably to be explained as a function of the inaccuracy of our method of measuring the "intelligence" of these subjects

According to Gates: "Sex differences are not manifest and differences caused by social status, if they occur, are slight" This applies, of course, to her experiment with white children. No figures are given to substantiate these statements Our own results, however, indicate a rather pronounced sex difference, as shown in Table IV. The boys are superior to the girls only in the younger ages (where in most instances the number of cases is small) From seven to fourteen years, on the other hand, girls exceed the boys. This superiority is to be explained no doubt as a result of the more rapid maturing of the girls The female subjects, actually in a more advanced stage of development than males of the same age, would hence be capable of better performance in mental tasks

TABLE IV.—PER CENTS OF SUCCESSES OF NEGRO BOYS *vs* NEGRO GIRLS
(All Schools Combined)

Ages	3	4	5	6	7	8	9	10	11	12	13	14
Boys												
Per cent	00	17	30	34	34	40	50	54	56	56	63	60
N	4	6	9	15	18	13	18	19	12	17	19	18
Girls												
Per cent	00	08	24	28	42	45	53	61	59	68	63	62
N	1	2	11	15	19	17	19	14	20	14	15	17

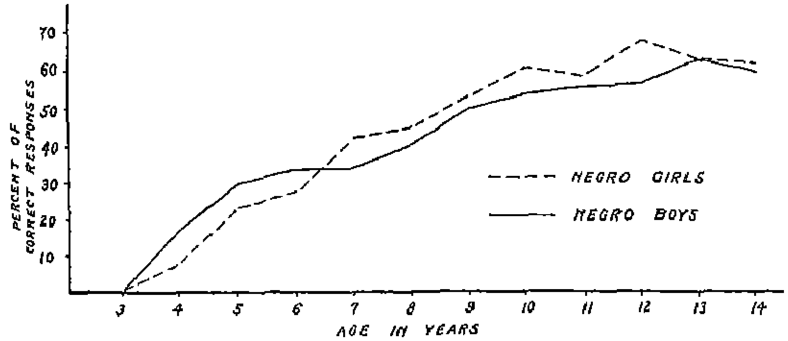


FIG 2 —Growth curves showing relative scores of negro girls and negro boys as given in Table IV Above age 6 the scores of the girls demonstrate a consistent tendency to be slightly higher than those of boys of the same ages

Differences between the major groups of the three public schools are indicated in Table V Two rather striking facts appear from the data in this table: (1) The subjects from the Bloomington school are superior to those of either of the other schools in five of the possible nine age-groups, (2) the subjects from Public School No 42 (presumably from average to superior social levels) exceed the other two groups at only one age

A comparison of the two Indianapolis schools, year by year, discloses that neither is consistently superior to the other The subjects from these two schools were specifically selected, however, because of differences in environmental levels It is indicated, therefore, from these figures that distinctions in social strata are of minor significance in the development of social perception The general superiority of the Bloomington children over the other groups can be accounted for by the fact that in this small school two or three grades are kept together in a single classroom. The continual contact with those older than themselves, affords the children greater opportunity to become adept in social perception than would be possible if they were segregated in separate classes.

TABLE V.—PER CENTS OF CORRECT RESPONSES FOR DIFFERENT SOCIAL GROUPS

Ages	6	7	8	9	10	11	12	13	14
Bloomington									
Per cent .	29	35	48	50	50	.65	75	68	69
N	7	9	10	12	11	11	8	12	13
Indianapolis public school No 42									
Per cent	.33	39	.42	51	55	.50	.63	60	.63
N	11	16	8	13	10	10	10	10	10
Indianapolis public school No 24									
Per Cent	31	40	30	53	64	58	51	61	50
N . .	12	12	12	12	12	11	13	12	12

IV SUMMARY AND CONCLUSIONS

Following the procedure adopted by G S. Gates for measuring the growth of social perception in white children, three hundred thirty-two negroes varying in age from three to fourteen years were similarly tested. Six pictures of the Ruckmick series were presented one at a time to each subject who then attempted to indicate the emotion represented.

1. Regardless of geographical differences in the location of the whites and negroes under examination and the fact that children of both stocks judged facial expressions posed by a *white* woman, a striking similarity was evidenced year by year in the data for the two racial groups.

2 The growth of social perception in negro children was hence demonstrated to be the same in all major respects as that which Gates found for white subjects

3. The negro girls were consistently superior to the negro boys except at the very young ages

4 A rough positive relationship appears between social perception as measured in this experiment and teachers' ratings of intelligence.

5 Groups of different social status manifest no superiority in the interpretation of facial expressions which can not be explained by factors other than their respective social levels

ON KIN RESEMBLANCES IN PHYSIQUE *VS.* INTELLIGENCE

HERBERT S. CONRAD*

Institute of Child Welfare, University of California

A number of studies have pointed out the similarity of familial coefficients of resemblance in physical and in mental traits. The inference, by analogy, is usually that the mental traits are inherited in the same way and to the same extent as the physical.[4] As a possible objection to this, Spearman long ago pointed out that the physical coefficients of resemblance are relatively unattenuated by the unreliability of measurement; whereas the mental coefficients are probably considerably attenuated.[13] One of our leading investigators in this field has recently found the resemblance of siblings in intelligence to equal a correlation of .60 (corrected for attenuation). He concludes:[17]

> If we may accept Pearson's results for the resemblance of siblings in eye color, hair color, and cephalic index (.52, .55, and .40), and regard .52 ± .016 as the resemblance in traits entirely free from environmental influence, we may infer that *the influence upon intelligence of such similarity in environment as is caused by being siblings two to four years apart in age in an American family to-day is to raise the correlation from .52 to .60.*†

This conclusion, while possibly correct, can hardly be accepted solely on the evidence presented by its author. The measure of intelligence employed, consisted of "a selection of tests from standard instruments—the Institute of Educational Research Tests of Selective and Relational Thinking, Generalization, and Organization."[18] Now, if the author had taken a *single test* from his total intelligence battery, and compared the correlation of siblings in *this test* with the correlation of siblings in (say) eye color, the comparison might have been acceptable. Eye color is, apparently, a relatively simple trait, determined by a small number of genes, and perhaps the score in a *single test* approaches genetic‡ comparability; but the *total* score in an entire intelligence test is almost certainly too complex, for a comparison with eye color to be genetically significant. Most students of the constitution of mental traits, from Spearman to Thomson to Kelley, appear agreed on at least one point: that the total score on an intelligence test represents a composite (with unknown weights) of several more-or-less intellectual abilities, or traits. The statistical effect of merging several component traits into a general, composite trait like "total intelligence-test score" is, in general, to increase—and, so far as the comparison with simple physical traits is concerned, to *spuriously* increase—the correlation between siblings in the general trait. A precise evaluation of the spurious rise in correlation may be secured by applying Spearman's formula for the correlation of sums and averages.[6]

---

* The manuscript has benefited from the criticism of Dr. Robert C. Tryon, National Research Fellow at the University of California.

† Italics as in the original.

‡ Throughout this paper, the word "genetic" is used in the biological sense of "hereditary," or "pertaining to genes," never in the sense of "pertaining to development," as in the term "genetic psychology."

Let us, for example, suppose that the average correlation between siblings in each of eleven separate tests is .52,* that is to say,

$$r_{1\mathrm{I}} = .52,\ r_{2\mathrm{II}} = .52,\ r_{3\mathrm{III}} = .52,\ \cdots\ r_{11\mathrm{XI}} = .52$$

This coefficient coincides with the average found by Pearson for certain physical traits (eye color, hair color, and cephalic index). Now let us add the eleven tests to form a single battery or "intelligence test." What is the correlation between the siblings in this total battery? If the average intercorrelation between the separate tests (for the identical individuals tested on the several tests) be .60, and if the average of the "cross-correlations" be .37—then, applying Spearman's formula, we find that the correlation between the siblings *in the total test* is .60 (see Table I). Would this increase from .52 to .60 represent, as appears to have been argued, the effect of environment, or would it rather represent a necessary statistical outcome of the combination of individual tests into a composite total score? Obviously, the coefficient of kin resemblance which one obtains for "intelligence" *depends to a marked extent on the number of tests which*

---

* Throughout this illustration, all *r*'s as given are assumed to be freed from the attenuation due to unreliability of measurement.[13]

The notation which follows is somewhat similar to that of Kelley.[7] *1* refers to Test 1 administered to the persons in the *X*-distribution; I refers to the same test administered to the *siblings* of the persons in the *X*-distribution; $r_{1\mathrm{I}}$ is therefore the correlation between siblings in Test 1. Similarly, $r_{2\mathrm{II}}$ is the correlation between siblings in Test 2, etc. In like fashion, $r_{1\mathrm{II}}$ would mean the "cross-correlation" between the scores of persons in Test 1, and the scores of their siblings in Test 2; $r_{2\mathrm{III}}$ would have a similar meaning for the sibling cross-correlation in Tests 2 and 3, etc. See Table I.

are merged into the total "intelligence" score, and the intercorrelation of these tests. With a suitable battery of tests called "intelligence," it seems that almost any degree of sibling resemblance might be obtained.*

TABLE I—THE CORRELATION BETWEEN SIBLINGS IN A TOTAL INTELLIGENCE TEST, DETERMINED FROM THEIR CORRELATION IN THE SEPARATE SUBTESTS

$$r_{(1+2+\cdots 11)(\mathrm{I}+\mathrm{II}+\cdots\mathrm{XI})} = \frac{ab\bar{r}_{pQ}}{\sqrt{a+(a^2-a)\bar{r}_{pq}}\sqrt{b+(b^2-b)\bar{r}_{PQ}}}\dagger$$

$(1+2+\cdots 11)$ means the sum of the scores in the eleven tests, made by the persons in the X-distribution

$(\mathrm{I}+\mathrm{II}+\cdots\mathrm{XI})$ means the sum of the scores in the same eleven tests, made by the *sibs* of the persons in X-distribution

$a = b =$ number of tests $= 11$

$\bar{r}_{pq} =$ the average intercorrelation between the tests, for the persons in the X-distribution $= 60$

$\bar{r}_{PQ} =$ the average intercorrelation between the tests, for the persons in the Y-distribution $= 60$

$\bar{r}_{pq}$ is assumed to equal $\bar{r}_{PQ}$, in this case, because the same tests are involved, only the sample receiving the tests differs nominally. (The sample from which $\bar{r}_{PQ}$ is obtained consists of the siblings of the sample from which $\bar{r}_{pq}$ is obtained.)

$\bar{r}_{pQ} =$ the average of: $r_{1\mathrm{I}} + r_{1\mathrm{II}} + r_{1\mathrm{III}} + \cdots + r_{1\mathrm{XI}} + r_{2\mathrm{I}} + r_{2\mathrm{II}} + r_{2\mathrm{III}} + \cdots r_{2\mathrm{XI}} + r_{3\mathrm{I}} + r_{3\mathrm{II}} + r_{3\mathrm{III}} + \cdots + r_{3\mathrm{XI}} + \cdots + r_{11\mathrm{I}} + r_{11\mathrm{II}} + r_{11\mathrm{III}} + \cdots r_{11\mathrm{XI}}$

(There are, in all, 11 coefficients of correlation with subscripts like *1*I, *2*II, *3*III, *4*IV, etc., each of these correlations equals 52. There are 110 coefficients of correlation with subscripts like *1*II, *1*III, . . . *1*XI, *2*I, *2*III, . . . *2*XI, *3*I, *3*II, *3*IV, . . . *3*XI, etc , the average of these 110 "cross-correlations," by the assumptions of this illustration, equals .37)

$$r_{(1+2+\cdots 11)(\mathrm{I}+\mathrm{II}+\cdots\mathrm{XI})} = \frac{(11)(11)\left[\frac{11(52)+110(37)}{121}\right]}{\sqrt{11+(121-11)(.60)}\sqrt{11+(121-11)(60)}}$$
$$= 60$$

---

* Some illustrations of this from experimental studies are available. Starch,[15] for example, finds the average correlation between eighteen pairs of adult siblings in ten separate achievement tests to be 42; in five separate mental tests, the correlation is .38; but in all fifteen tests combined, the correlation is 73! Somewhat similarly, Jones[5] finds the correlation between single parent and single child to equal .548; but the correlation of mid-parent with mid-child (average score of all the children) is .651. Willoughby,[20] making use of scores in *single tests* only, and correcting for attenuation, reports an average sibling coefficient of .42; this, as Terman remarks, is "out of line with (lower than) the results of other investigators."[16] These "other investigators" have, in general, used *total* intelligence test scores.

† *Cf* Reference 6

It is evident, then, that to measure a functional complex, like "intelligence," and to assume this complex as a genetic entity,* may involve statistical difficulties Of course it is impossible, in the realm of mental characters, to measure the strictly anatomical structures to which the genetic entities most probably correspond The point here is simply that the (different) ability complexes which are measured by (different) intelligence tests are not, in any likelihood, the *functional counterparts* of genetic entities *"Intelligence," as measured, demands analysis* From this point of view, the mathematical studies of Rosenow,[11] Spearman,[14] Hull,[3] and Kelley[8] would appear of a fundamental nature, and not as *objets d'art* [12] It is, however, unfortunate that these useful and stimulating mathematical contributions have not more specifically considered genetic problems and mechanisms

Another statistical misunderstanding deserves mention. In reviewing the literature one occasionally finds the statement, or somehow secures the impression, that because the coefficient of correlation between kin is 20 (or 30 or 40, etc), the evidence for kin-resemblance is strong In truth, however, a correlation of .20 or 30 or .40 is indicative not of strong similarity, but of fairly strong dissimilarity A more legitimate assertion from these correlations, perhaps, is that the evidence for heredity is significant. As a matter of fact, however, at least in the realm of mental traits, nobody knows (except by doubtful analogy with physical traits) what correlation should be expected between kin, if heredity alone were the cause of resemblance. Nobody knows, because nobody yet knows (except very hypothetically) the mechanism of mental inheritance † Serious argument from empirical correlation coefficients *alone*, therefore, seems as yet unjustified. To give extreme examples for the emphasis of this point: The correlation between parent and offspring for trait X, in a pure, inbred, homozygous line, would (if the parental environmental factors are uncorrelated with the offspring's) be zero.[22] And yet this absence of correlation is consistent with a quite profound influence of heredity, which, indeed, could never be deduced from the zero correlation coefficient *alone*, even if statistical allowance were made for the restriction of range in an inbred population.[7a, *] Conversely, the correlation between tuberculous infection in parents and tuberculous infection in offspring may be quite high; but this high correlation is consistent with a negligible or even zero influence of heredity (in the strict sense of the word). Perhaps the most embarrassing situation arises when the correlation which might be expected through the exclusive influence of heredity, coincides with that which might be expected through the exclusive influence of environment.

---

* By "genetic entity" the writer means a *single elementary trait* which is caused, in part or in whole, by a specific genetic mechanism—whether unit-factor or multiple-factor Thus, *white forelock* is a genetic entity (probably unit-factor), but *handsomeness* is not a genetic entity, because (like intelligence-test score) it is a *complex* trait composed of *several* individual (and arbitrarily weighted) entities

† R A Fisher[1] and Sewall Wright[21] have postulated certain coefficients of phenotypic resemblance between kin, on the basis of certain assumptions concerning the mechanism of heredity Some of Fisher's assumptions (particularly as to the number of genetic factors involved) may be doubtful in themselves, and Fisher himself admits that his whole set of assumptions may be over-simple (such possible phenomena as lethal factors, somatic mutation, environmental effect upon dominance, differential fertility of germs, sexual selection, and the correlation of hereditary and environmental factors are ignored) Wright gives the most complete set of predictions known to the writer, but these predictions are based upon definite

A great deal of intrinsic interest attaches to the relative magnitude of kin resemblances in intelligence, height, cephalic index, eye color, etc. The point seems to deserve emphasis, however, that merely biometric studies of kin resemblance (under uncontrolled conditions of environment), do not yet have the necessary psychological, genetic, *and mathematical basis for anything like rigorous, quantitative* interpretation of results. Experimental or pre-experimental studies of the type of Burks,[23] Freeman,[23] Muller,[9] Newman,[10] Gesell,[2] and Tryon[10] seem indicated.

NOTE.—Discussion of the foregoing paper with Dr. Tryon has brought forth the following points.

1. The rise in correlation attending the pooling of tests may not be spurious, in the sense that the rise may be interpreted in two ways. (*a*) One possible cause of the rise may be the closer balancing, in the pool, of factors uncorrelated between the siblings (this accords with Spearman's two-factor theory of mental abilities). Whether either the general factor or the specific factors are hereditary or environmental or both, is not divulged by the correlation coefficients alone. (*b*) An alternative cause of the rise may be the fuller sampling of abilities in the pool than in any individual test (this accords with a multiple-factor theory).

---

assumptions concerning both the mode of inheritance and the system of mating. In the case of human mental inheritance, ignorance of the mode of inheritance would appear to preclude confident use of any one of Wright's formulæ.

[*] Inasmuch as a zero coefficient remains unaltered by the correction for restricted range.[7a]

2 A difference between parent-child correlations in two psychologically unitary traits may argue as well for a different mode of inheritance as for a differential effect of environment

3. In the comparison of correlation coefficients, due consideration must be given to the nature or make-up of the variables concerned It is worth noting that the effect of pooling usually diminishes rapidly, as the size of the pool increases This suggests the desirability of comparing either simple elementary traits, in which the effect of pooling is absent; or highly complex traits, in which the effect of pooling is fairly stable, and perhaps uniform

REFERENCES

1 Fisher, R A The Correlation between Relatives on the Supposition of Mendelian Inheritance *Transactions Royal Society Edinburgh*, 1918, Vol II, pp 399–433 (see especially pp 400, 401, 420, 432)

2 Gesell, A, and Thompson, H. Learning and Growth in Identical Infant Twins *Genetic Psychology Monograph*, 1929, No 6

3 Hull, C. "Aptitude Testing," Chaps 6, 7 World Book Co, Yonkers, N Y, 1928

4 Jones, H E A First Study of Parent-Child Resemblance in Intelligence The *Twenty-seventh Yearbook of the National Society for the Study of Education*, 1928, Vol I, p 72

5 *Ibid* P 69.

6 Kelley, T L "Statistical Method" P 204, formula 154 The Macmillan Company, 1924

7. *Ibid* Pp 196–197.

7a *Ibid* P 225, formula 187

8 Kelley, T L "Crossroads in the Mind of Man" Stanford University Press, 1928

9 Muller, H J Mental Traits and Heredity *Journal of Heredity*, 1925, Vol XVI, pp 433–448

10 Newman, H H Mental and Physical Traits of Identical Twins Reared Apart *Journal of Heredity*, 1929, Vol XX, pp 49–64, 97–104, 153–166

11 Rosenow, C The Analysis of Mental Functions *Psychology Monograph*, 1917, Vol XXIV, No 5

12 Slocombe, C Truman L Kelley Measures Mental Traits *Journal of Educational Psychology*, 1928, Vol XIX, p 501

13 Spearman, C The Proof and Measurement of Association between Two Things *American Journal of Psychology*, 1904, Vol XV, pp 72–101

14 Spearman, C "The Abilities of Man" The Macmillan Co., N Y, 1927

15 Starch, D Similarities of Brothers and Sisters in Mental Traits *Psychological Review*, 1917, Vol XXIV, pp 235–238

16 Terman, L M Nature and Nurture Their Influence upon Intelligence and upon Achievement *Journal of Educational Psychology*, 1928, Vol XIX, p 367

17 Thorndike, E L The Resemblance of Siblings in Intelligence The *Twenty-seventh Yearbook of the National Society for the Study of Education*, 1928, Vol I, pp 41–53

18. *Ibid.:* P. 41.
19. Tryon, R. C.: The Genetics of Learning Ability in Rats. *University of California Publications in Psychology*, 1929, Vol. IV, pp. 71–89.
20. Willoughby, R. R.: Family Similarities in Mental-test Abilities. The *Twenty-seventh Yearbook of the National Society for the Study of Education*, 1928, Vol. I, pp. 55–59.
21. Wright, S.: Systems of Mating, I. The Biometric Relations between Parent and Offspring. *Genetics*, 1921, Vol. VI, pp. 111–123.
22. Wright, S.: Systems of Mating, V. General Considerations. *Genetics*, 1921, Vol. VI, p. 169.
23. In the *Twenty-seventh Yearbook of the National Society for the Study of Education*, 1928, Vol. I.

A NOTE ON METHODS OF MEASURING RELIABILITY

T G FORAN

Catholic University of America

Three procedures are available for measuring the reliability of tests According to the first, the same form of the test is administered twice The second involves estimating the reliability of the entire test or scale from the correlation of its halves. In the third method two forms of the test, each given once, furnish the data from which the measures of reliability are calculated. The third procedure is generally preferred to the others All methods involve obvious differences under which the reliability of a test is found.

The data in this study were compiled for the purposes of comparing the first and third methods as described above Four schools were used In one, the first list of the Morrison-McCall Spelling Scale was given twice to all pupils from the Grades II to VIII, inclusive. In the second school, the second list of the scale was used twice In the third school, the first and second lists were used in that order In the fourth, the second and first lists were used In all grades of the four schools, a day intervened between the first and second tests. The rules for administering and scoring the tests stated in the scale were followed precisely and all computations were performed twice on a calculating machine.

Identifying information, means, standard deviations, and several measures of reliability are presented in Table I. The averages of the reliability coefficients are higher when the same form is repeated than when duplicate forms are employed All reliability coefficients with one exception are higher when either List 1 or List 2 is repeated than when the two lists are employed and comparisons are made between the same grades in the four schools Nine of the twelve probable errors of score are lower for the single form method Further confirmation of these results is obtained from the probable errors of estimated true scores

In order to render the reliability coefficients directly comparable through holding the variability constant, each coefficient has been used with Kelley's formula·

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1 - R}}{\sqrt{1 - r}}$$

with Σ or the larger standard deviation arbitrarily set at 10. In nine of the twelve possible comparisons of the same grades in the four schools, higher coefficients are secured from the two applications of the same form. In one of the three comparisons in which the reverse occurred, the difference is negligible.

In Table II the measures of reliability have been averaged. With three grade-groups and two schools for each method, six measures contribute to each of the averages. The difference between the means of the coefficients of reliability appears to be considerably greater than between the *R*'s. A better comparison is possible through the measures of improvement over chance.

	Same form twice	Duplicate forms	Difference
r	935	866	
I_p	64 5	50 0	14 5
R	969	951	
I_n	75 3	69 1	6 2

TABLE I.—MEASURES OF RELIABILITY OBTAINED WITH DIFFERENT ARRANGEMENTS OF TESTS

Order of tests	Grades	N	M_a	M_b	SD_a	SD_b	r		R[1]	PE_{score}[2]	$PE_{est\ t\ s}$[3]
1-1	2, 3, 4	95	19 12	20 60	6 51	6 08	911	012	961	1 33	1 27
1-1	5, 6	66	32 06	35 51	8 20	7 50	930	011	956	1 41	1 30
1-1	7, 8	60	13 30	11 01	1 77	4 30	933	011	980	70	76
Mean							925		968	1 18	1 13
2-2	2, 3, 4	80	20 08	21 82	9 00	7 60	972	005	985	82	81
2-2	5, 6	56	32 60	35 20	8 07	7 51	897	017	938	1 60	1 60
2-2	7, 8	53	10 06	11 07	6 22	5 13	968	005	980	70	69
Mean							910		971	1 07	1 03
1-2	2, 3, 4	213	18 93	19 78	7 39	6 71	893	008	947	1 56	1 47
1-2	5, 6	152	32 21	31 24	5 90	6 10	883	012	957	1 30	1 31
1-2	7, 8	110	11 37	10 19	4 51	5 01	838	010	963	1 30	1 19
Mean							871		950	1 12	1 32
2-1	2, 3, 4	125	18 86	20 18	8 38	8 88	926	009	945	1 58	1 52
2-1	5, 6	100	33 66	35 19	7 30	6 15	858	017	948	1 53	1 41
2-1	7, 8	91	11 60	10 98	4 62	5 15	707	020	944	1 50	1 39
Mean							860		946	1 57	1 45

---

[1] R as obtained from Kelley's formula $\frac{\sigma}{\Sigma} = \frac{\sqrt{1 - R}}{\sqrt{1 - r}}$ with $\Sigma = 10$

[2] $PE_{score} = 6745\,\sigma\sqrt{1 - r}$

[3] $PE_{est\ t\ s}$ = probable error of estimated true score = $6745\sigma\sqrt{r - r^2}$

TABLE II —MEANS OF SIX MEASURES OF RELIABILITY FOR EACH ARRANGEMENT OF TESTS

	Same form twice (1-1, 2-2)	Duplicate forms (1-2, 2-1)
r	935	866
R	969	951
PE_{score}	1 12	1 49
$PE_{est\ t\ s}$	1 08	1 39
Total number of pupils	416	869

The improvement over chance has been computed from the formula [1]

$$I_p = 100(1 - \sqrt{1 - r^2})$$

While the means of the reliability coefficients over-emphasize the difference between the methods, a substantial difference remains when variability is held constant

Table III contains the significance ratios for the combinations of reliability coefficients The significance ratio is the difference between the coefficients divided by the probable error of the difference. The probable errors of the difference have been found by means of the usual formula [2]

TABLE III —SIGNIFICANCE RATIOS OF DIFFERENCES BETWEEN RELIABILITY COEFFICIENTS

Tests	Grades			Mean
	II, III, IV	V, VI	VII, VIII	
1-1 and 2-2	−4 69	+1 65	−2 92	3 09
1-2 and 2-1	−2 75	−0 24	+2 15	1 71
1-1 and 1-2	+1 29	+2 94	+5 00	3 08
1-1 and 2-1	−1 00	+2 10	+5 35	2 82
2-2 and 1-2	+8 78	+0 67	+7 04	5 70
2-2 and 2-1	+4 60	+0 38	+6 93	3 97

$$PE_{diff} = \sqrt{PE^2_{r-1} + PE^2_{r-2}}$$

---

[1] Holzinger, Karl J "Statistical Methods for Students in Education " Ginn and Co , New York, 1928, p 166

[2] Garrett, Henry E "Statistics in Psychology and Education " Longmans, Green and Co , New York, 1926, p 171

The significance ratios are positive when the first coefficient of the pair is higher than the second. The means of the ratios have been found without regard to the signs of the ratios. It is noteworthy that the order of the tests has no influence on the results when the tests are similar forms. The most important differences occur when a reliability coefficient obtained from the same form used twice is compared with the reliability coefficient from similar forms. The significance ratios are much larger for Grades VII and VIII combined than for Grades V and VI together. In general, reliability coefficients from repeated tests are significantly higher than the same measures found from one use of each form of the test.

TABLE IV.—PRACTICE EFFECTS AND EQUIVALENCE OF LISTS OF MORRISON-McCALL SPELLING SCALE

Grades	Order of tests	*N*	Difference	
			+	−
II, III, IV	1-1	95	1.48	
V, VI	1-1	66	2.85	
VII, VIII	1-1	66	1.61	
Mean	1-1		1.98	
II, III, IV	2-2	80	.84	
V, VI	2-2	50	2.60	
VII, VIII	2-2	53	1.01	
Mean	2-2		1.48	
II, III, IV	1-2	243	.85	
V, VI	1-2	152		1.00
VII, VIII	1-2	149		1.18
Mean	1-2			.44
II, III, IV	2-1	125	1.32	
V, VI	2-1	109	1.83	
VII, VIII	2-1	91		.71
Mean	2-1		.81	

Difference is positive when second test has higher mean. Means are not weighted.

Some measure of practice effect under the same conditions is possible from the data at hand. Table IV contains the differences between the means for all grades and schools. The practice effect involved in taking the same test twice is considerable. It is greater for List 1 than for List 2, the means being 1 98 and 1.48 but this may be a sampling error The algebraic mean of the differences between the two tests in the last six groups is + 19. There are some indications that List 2 is more difficult than List 1. When the order of the tests is 1-2, the mean gain is 44 but when the order is reversed the mean gain is .81

CONCLUSIONS

1. When allowances are made for differences in variability, reliability coefficients are higher for repetitions of the test than for similar forms.

2 The order in which the tests are given has no effect on the results, even when one form is slightly more difficult than the other

3. The practice effect of taking the same test twice is estimated to be 1 5 times the practice effect from duplicate forms This obtains only for the Morrison-McCall Spelling Scale

4. In view of the results of this study, it is important to consider the conditions underlying the determination of test reliability

5 Using the method of two forms, the probable error of estimated true scores of the Morrison-McCall Spelling Scale (and probably for any well scaled spelling test of fifty words) is 1 4 words.

SAMPLING ERROR OF TETRAD DIFFERENCES

C. SPEARMAN

University of London

In the now prevalent usage of the formula for the probable error or variance of tetrad differences,[1] there is apt to be scant account taken of the special assumptions upon which this formula is based. One of these, namely that the variables should have a normal frequency distribution, is only dangerous when great accuracy is wanted.[2] But the other assumption can do much more harm; it is that the correlations are measured by the ordinary product moment coefficients, so that the variance of the coefficient

$$\sigma_r^2 = (1 - r^2)^2/N.$$

One case where this latter assumption may be far from justified is that of coefficients which have been corrected for attenuation. And even worse is the case of coefficients derived from any of the customary four-fold tables. On the other hand, the coefficients derived from ranks (with the differences squared) would not seem to introduce any disturbance of appreciable size.

Of course, there is nothing to prevent anyone from so modifying the formula for the tetrad sampling error that it does become valid for any coefficient required. All that has to be done is that in deriving the formula the above quoted σ_r^2 should be replaced by the variance which is appropriate to that coefficient. Such variances can be found in any good textbook; for example, in those of Holzinger, Garrett, or Kelley

---

[1] See the present writer's "Abilities of Man," appendix, pp x–xi (except that on p. x, N should be replaced by $N^{1/2}$)

[2] See this JOURNAL, Nov., 1930, Disturbers of Tetrad Differences

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

CONDUCTED BY FRANCES M. FOSTER

Abnormal Psychology, Its Concepts and Theories, by H L Hollingworth New York The Ronald Press Company, 1930 Pp XI + 590.

A very distinctive as well as significant contribution to the field of abnormal psychology is Dr H L Hollingworth's latest book. This work represents a serious effort to make the concepts and theories concerning mental abnormalities psychologically intelligible It also represents a determined effort to systematize these concepts and theories The system constructed has a logical place in it for facts of consciousness along with behavior but the evaluations and reinterpretations of materials selected for presentation can hardly be said to be characterized by catholicity. Most viewpoints are evaluated but the psychoanalytic—labelled by Hollingworth, the psychoanalogical—is vigorously attacked. His own conceptual system is consistently applied. In this system the general pattern of mental activity, abnormal as well as normal, is said to be characterized by the redintegrative sequence "Partial stimuli now occurring function for former antecedents of greater complexity." Learning and sagacity are the two important aspects of these processes "Without learning there could be no mental activity, without sagacity, mental activity becomes biologically and socially ineffective." Feeblemindedness is, accordingly "relative inaptitude for redintegrative activity." "Unsagacious redintegrated responses give the picture of the psychoneuroses On the postural level such symptoms give the somatic, physical picture of hysteria; on the autonomic level, they comprise the complaints of the neurasthenic, on the symbolic level they show themselves in fixed ideas, obsessive thoughts, morbid pictures and ruminations, as in psychasthenia." Psychotherapy consists in teaching the patient to substitute a symbolic for a postural or an autonomic response. Lack of space prohibits a fuller consideration of the fundamental conceptions in this system They are the same concepts which the author has previously used in reconstructing the facts and principles in general psychology. And the reconstruction is practically as thoroughgoing.

In the Preface the author differentiates his own interest in mental abnormality as a student of natural phenomena from the "everyday" interests of educators or hygienists and the "technological" interests of psychiatrists or clinical psychologists. This implies to him a dominant interest in concepts rather than clinical pictures, ideas rather than cases, principles rather than procedures; in short, an interest in understanding rather than preventing or relieving mental abnormality. Many cases are presented—more than in some of the books which emphasize clinical pictures. The book really contains more practical features than the foregoing statements imply. The chief reason, of course, is that in dealing with problems of human behavior the point of demarcation between diagnosis and treatment is hardly perceptible—the goal from the beginning to end can be said to be fuller understanding.

The order of presentation as well as the organization of the materials in each of the twenty-five chapters is apparently the result of much conscious planning The outline is clear and the march continous. The book begins with a succinct chapter on the subject-matter and uses of abnormal psychology. Abnormal behavior is here differentiated from desirable behavior. "It may be comfortable," we are told, "to be normal, since this will mean that we are not conspicuous, but normality, in the scientific sense, is nothing on which to pride oneself"

Following the introductory chapter are six very comprehensive chapters describing historical and contemporary viewpoints. The only one that can not be passed by without comment is the one which purports to give a fair account of psychoanalysis. There is no denying the validity of a good deal of the criticism Nor can one deny Hollingworth's greater clarity. But one is tempted to add that at least some of the clarity is obtained by ignoring fundamental difficulties which confront psychologists who are more concerned with people rather than concepts

In the consideration of personality deviations the author first attacks the simpler and more measurable deviations, advances to the "Dwellers of Neurotica" and thence to the psychotics and other more complex deviations. Two chapters are given to the feebleminded, nine to the psychoneuroses and one chapter a piece to each of the following topics: Stage-fright and dream, stuttering and stammering;

aphasia and asymbolia, epilepsy, constitutional psychopathic states, personality types and functional psychoses; and mental disorder and the effect of drugs

Objective studies of the neurotic and the feebleminded are fully treated In the discussion of the neuroses more than the usual amount of attention is devoted to the contributions of Babbinski, Hurst and Rosanoff as well as those of Janet and Prince The psychoanalytic contributions of Freud, Adler and Rivers are presented as Herbartian conceptions of the neuroses. The result is interesting reading but can hardly be considered as more than a forced interpretation, a waste-basket variety of categorizing Hollingworth's explanation of primitive behavior in terms of redintegrative sequence is no more realistic than Freud's which he criticizes—neither is based on actual anthropological studies No more convincing is the description of all Freudian forgetting as "underlearned material"

A critical evaluation of fictional neuro-anatomy, a valid criticism of Cotton's overvaluation of the rôle of infection in mental disorders, a much needed emphasis on the rôle of the learning process in the development of the functional disorders, a healthy emphasis on the how as well as the what of the thinking of the contributions considered—these, in the opinion of the reviewer, are among the most pleasing as well as distinguishing features of the book

H MELTZER.

Psychiatric Clinic, Saint Louis, Missouri.

The Guidance of Mental Growth in Infant and Child, by Arnold Gesell. New York. The Macmillan Company, 1930 Pp XI + 322

If parents and educators more frequently combined their avowed love and respect for little children with the sympathetic point of view gained by scientific study of mental development, many child adjustment problems would be solved. Dr Gesell shows a rare appreciation and respect for childhood's complexities, and he beholds with wonderment the orderly progression of mental unfoldment which his own techniques, skillfully applied, have revealed He shatters many an illusion and yet builds up our faith in the possibilities of solving behavior problems. If in his enthusiasm the author verges at times on the sentimental and stresses the obvious, these tendencies do not detract from the soundness of the principles discussed There is occasional repetition due to inclusion of previously published articles which overlap in content.

Children have not always been so sympathetically understood, but the growing tendency toward clearer conceptions of the purpose of childhood and parent-child relationships is revealed in the accounts of early nursery school projects and in the correspondence of intelligent mothers of whom Susanna Wesley was an illustration Older concepts of child guidance are illustrated with an amusing and charming series of old prints and lithographs The modern nursery school with its attendant program of scientific child study is the newest addition to institutional education and promises to integrate the activities of home and school of children below kindergarten age. Instead of the kindergarten giving formal instruction based on dubious psychological principles, we find the modern kindergarten assuming an important rôle in child guidance, forming a continuum with the nursery school and the elementary grades.

Concerning the mental development of the child as determined by the ingenious methods of the Yale Psycho-clinic, the author has many significant findings to report The most important principles formulated are those concerning the rapidity, complexity and orderliness of mental growth from birth to maturity.

Dr Gesell formulates normality of mind in terms of. (1) Wholesome personal habits of living; (2) wholesome habits of feeling; and (3) healthy attitudes of action

Concerning the part played in development of the two factors of heredity and environment there has been a great deal of misunderstanding Dr. Gesell's treatment of the topic is one of the sanest and most interesting of existing reports He finds that the child is remarkably well insulated against chance environmental influences and conditioning and that much of what in the young child appears to be the result of training and teaching is really the result of mental unfoldment and natural maturation. Because the child is a human being certain behavioristic trends are bound to assert themselves in spite of even decidedly adverse environmental conditions Physical handicaps do not drastically alter the behavior capacities of the child. The inevitableness and surety of maturation safeguards the child against adventitious circumstances

The validity of the conclusions reported is well established and their application in child guidance work will be unquestionably beneficial Every one who has child training responsibilities will find Dr Gesell's book interesting and valuable.

Gertrude Hildreth

The Lincoln School of Teachers College, Columbia University.

THE REVIEW OF EDUCATIONAL RESEARCH

The first issue of the first journal to attempt a systematic review of educational research has just appeared The Review is the official organ of the American Educational Research Association which is now a Department of the National Education Association.

The first issue deals with the curriculum It was prepared by a committee consisting of Dr. Henry Harap, Dr. William L Connor, and Dr Ralph W. Tyler, assisted by several other persons. This issue includes an extensive classified bibliography and treats of methods of curriculum-making, studies of the objectives of the curriculum, studies of learning which bear upon the curriculum, time allotment and grade placement, etc. Four other issues dealing with teacher personnel, school organization, special methods up to the end of the elementary school, and individual differences and psychological tests, are scheduled to appear during the year The whole field of research has been divided into fifteen topics which will be covered in three years. Three numbers will appear during the spring and two during the fall.

PUBLICATIONS RECEIVED

EDUCATIONAL PSYCHOLOGY

GENERAL

BOLTON, FREDERICK ELMER. *Adolescent Education* New York: The Macmillan Co., 1931, pp. XV + 506.

CRAWFORD, CLAUDE C. and LEITZELL, EDNA MABLE. *Learning a New Language* Los Angeles: University of Southern California, 1930, pp. XII + 242.

EZEKIEL, MORDECAI: *Methods of Correlation Analysis* New York. John Wiley and Sons, Inc., 1930, pp XIV + 427

GATES, ARTHUR I.. *Interest and Ability in Reading* New York: The Macmillan Co., 1930, pp XII + 264.

FILTER, RAYMOND O. and HELD, OMAR C.: *The Growth of Ability* Baltimore: Warwick and York, Inc., 1930, pp VII + 174.

ISAACS, SUSAN: *Intellectual Growth in Young Children.* New York: Harcourt, Brace and Co., 1930, pp. XI + 370

KATZ, DANIEL and ALLPORT, F. H.. *Students' Attitudes.* Syracuse, N. Y: The Craftsman Press, Inc., 1931, pp. XXVIII + 408.

PYLE, WM. HENRY: *The Psychology of the Common Branches* Baltimore. Warwick and York, Inc., 1930, pp. VII + 380.

ST. JOHN, CHARLES W.: *Educational Achievement in Relation to Intelligence.* Cambridge: Harvard University Press, 1930, pp XIV + 219.

STARCH, DANIEL: *Experiments and Exercises in Educational Psychology* New York. The Macmillan Co., 1930, pp. IX + 254.

WHEAT, HARRY GROVE: *The Psychology of the Elementary School.* New York: Silver, Burdett and Co., 1931, pp VII + 440.

THE CURRICULUM, TEACHING METHODS, AND THE PSYCHOLOGY OF LEARNING

EWING, IRENE R. *Lipreading.* Manchester Manchester University Press, 1930, pp. IX + 74

LINNELL, ADELAIDE: *The School Festival.* New York. Charles Scribner's Sons, 1931, pp. XXII + 124

LOWTH, FRANK J.: *The Country Teacher at Work* New York The Macmillan Co., 1930, pp. XII + 541

OSBURN, W J. and ROHAN, BEN J *Enriching the Curriculum for Gifted Children* New York. The Macmillan Co, 1931, pp XIV + 408

PETERS, CHARLES CLINTON *Objectives and Procedures in Civic Education* New York Longmans, Green and Co, 1930, pp VII + 302

ROCHELEAU, CORINNE and MACK, REBECCA *Those in the Dark Silence* Washington, D C. The Volta Bureau, 1930, pp. 169

TERRY, PAUL W., *Supervising Extra-curricular Activities* New York: McGraw-Hill Book Co, Inc, 1930, pp XI + 417.

TESTS, SCORE CARDS, TEACHERS GUIDES AND PUPILS GUIDES

BADANES, SAUL *Teacher's Book.* New York: The Macmillan Co., 1930, pp XIV + 101

CARRIGAN, ROSE A *Carrigan Score Card for Rating Teaching and the Teacher.* Yonkers-on-Hudson, N. Y · World Book Co, 1930.

GREENE, CHARLES E. and NOAR, FRANCES M. *Greene-Noar Self-diagnostic Reading Test.* New York· Heath and Co, 1931.

LEONARD, PAUL J · *Leonard Diagnostic Test in Punctuation and Capitalization.* Yonkers-on-Hudson, N. Y: World Book Co, 1931

LEONHARDY, ALMA: *Directed Study Guides for Stevenson's Treasure Island.* New York: The Macmillan Co., 1930.

WINSLOW, LEON LOYAL *Art Education Charts* Baltimore Warwick and York, Inc., 1930.

PROGNOSES, TESTING, AND MEASUREMENT

Educational Records Bulletin No. 5. *Testing School Achievement in England and America.* New York Educational Records Bureau, 1930

EDGERTON, HAROLD A.: *Academic Prognosis in the University* Baltimore· Warwick and York, Inc, 1930, pp VII + 83.

MADSEN, I N *Educational Measurement in the Elementary Grades* Yonkers-on-Hudson, N. Y. World Book Co, 1930, pp X + 294.

MANN, CLAIR V.· *Objective Type Tests in Engineering Education* New York: McGraw-Hill Book Co, Inc., 1930, pp. X + 121

MICHELL, ELENE *Teaching Values in New-type History Tests* Yonkers-on-Hudson, N Y.: World Book Co, 1930, pp IX + 179

PINTNER, RUDOLPH. *Intelligence Testing.* New York: Henry Holt and Co., 1931, pp XII + 555.

STUTSMAN, RACHEL *Mental Measurement of Preschool Children.* Yonkers-on-Hudson, N. Y.: World Book Co , 1931, pp X + 368.

WATERMAN, FLORENCE· *Studies and Tests on Vergil's Aeneid* Cambridge· Harvard University Press, 1930, Bulletin.

MONOGRAPHS

BECK, SAMUEL J.: *The Rorschach Test and Personality Diagnosis* Reprint, American Journal of Psychiatry, 1930.

BOHAN, JOHN E.: *Students' Marks in College Courses* Minneapolis: University of Minnesota Press, 1931.

BRANDENBURG, G. C. *Studies in Higher Education XVI* Lafayette, Indiana: Purdue University, 1930.

BUSWELL, G T. and JOHN, LENORE: *The Vocabulary of Arithmetic* Chicago: The University of Chicago, 1931.

CABOT, STEPHEN P.: *Secondary Education in Germany, France, England, and Denmark.* Cambridge Harvard University Press, 1930.

CARROLL, HERBERT ALLEN: *Generalization of Bright and Dull Children.* New York City: Bureau of Publications, Teachers College, Columbia University, 1930, pp VIII + 54

CASON, HULSEY: *Common Annoyances.* Princeton, N J. and Albany, N. Y.: Psychological Review Co , 1930.

CONRAD, HERBERT S. and HARRIS, DANIEL. *The Free-association Method and the Measurement of Adult Intelligence.* Berkeley, Calif. University of California Press, 1931.

HILGARD, ERNEST R.: *Conditioned Eyelid Reactions to a Light Stimulus Based on the Reflex Wink to Sound.* Princeton and Albany· Psychological Review Co , 1931.

JONES, CLINTON MELLEN: *Field Notes on Connecticut Birds.* Iowa City· University of Iowa City, 1931.

KELLOGG, W. N : *An Experimental Evaluation of Equality Judgments in Psychophysics* New York: Archives of Psychology, 1930.

KLOPP, WILLIAM J.: *The Relative Merits of Three Methods of Teaching General Science in the High School* Chicago: Central Association of Science and Mathematics Teachers, Inc., 1930.

LATON, AMITA DUNCAN. *The Psychology of Learning Applied to Health Education through Biology.* New York· Bureau of Publications, Teachers College, Columbia University, 1929, pp VI + 103.

Macfarlane, D. A. *The Rôle of Kinesthesis in Maze Learning.* Berkeley University of California Press, 1930.

Smith, Fred C.: *Curriculum Problems in Industrial Education.* Cambridge: Harvard University Press, 1930

Tolman, E C. and Honzik, C. H.: *Degrees of Hunger, Reward and Non-reward, and Maze Learning in Rats.* Berkeley: University of California Press, 1930.

Wood, Ernest Richard: *A Graphic Method of Obtaining Coefficients of Three or More Variables* Chicago: University of Chicago, 1931

FOREIGN PUBLICATIONS

Bertrand, Francois-Louis: *Alfred Binet et Son Oeuvre* Paris Librairie Felix Alcan, 1930

Bertrand, Francois-Louis: *L'Analyse Psycho-sensorielle.* Paris: Librairie Felix Alcan, 1930.

Bonaventura, Enzo: *Psicologia Dell' Eta Evolutiva.* Lanciano. Giuseppe Carabba, 1930.

Claparede, Dr Ed . *L'education Fonctionnelle* Paris: Delachaux and Niestle S.A., 1931.

Coade, T. F. (Editor): *Harrow Lectures on Education.* Cambridge. University Press, 1931, pp. XVII + 230

Croner, Else. *Doe Psyche der weiblichen Jugend.* Langensalza. Hermann Beyer and Sohne, 1930.

Firmo, Oliveira de Bruno, Annibal *O Exame Alpha.* Recife Diario da Manha, 1930.

Hermann, Dr. med *Krankhafte Seelenzustande beim Kinde.* Langensalza: Hermann Beyer and Sohne, 1930

Klages, L.. *Les Principes de la Caracterologie* Paris: Librairie Felix Alcan, 1930.

Koch, Dr. Hedwig *Das Generationsproblem in der deutschen Dichtung der Gegenwart.* Langensalza Hermann Beyer and Sohne, 1930

Kruisinga, E.. *An Introduction to the Study of English Sounds.* Over Den Dom-Utrecht: Kemink and Zoon N. V , 1931, pp XI + 168.

Pernambucano, Ulysses and Barretto, Annita Paes: *Estudo Psychotechnico de Alguns Tests de Aptidao.* Recife Imprensa Industrial, 1927.

Pernambucano, U. and Barretto, A. P *Ensaio de Applicacao do Test das* 100 *Questoes de Ballard.* Hungaro-Brasileira Ltd , 1930.

Pfahler, Dr. Gerhard: *Eros und Sexus.* Langensalza: Hermann Beyer and Sohne, 1930.

Schneider, Dr. Ernst: *Psychoanalyse und Padagogik.* Langensalza: Hermann Beyer and Sohne, 1930.

Ziehen, Th.: *Das Seelenleben der Jugendlichen.* Langensalza: Hermann Beyer and Sohne, 1931.

Psychology, Growth, and Development of Children

Buhler, Karl: *The Mental Development of the Child.* New York: Harcourt, Brace and Co., 1930, pp. XI + 170.

Eipper, Paul: *Human Children.* New York: Viking Press, 1930, pp. 70.

Gesell, Arnold: *The Guidance of Mental Growth in Infant and Child.* New York: Macmillan Co., 1930, pp. XI + 322.

Hartwell, S. M.: *Fifty-five "Bad" Boys.* New York: Alfred A. Knopf, 1931, pp. XVII + 359.

Porter, Martha Peck: *The Teacher in the New School.* Yonkers-on-Hudson, N. Y.: World Book Co., 1930, pp. XI + 312.

Rand, W., Sweeny, M. E., and Vincent, E. L.: *Growth and Development of the Young Child.* Philadelphia: W. B. Saunders Co., 1930, pp. 394.

Scharlieb, Mary: *The Psychology of Childhood Normal and Abnormal.* New York: Richard R. Smith, Inc., 1930, pp. XI + 194.

Stern, William: *Psychology of Early Childhood.* New York: Henry Holt and Co., 1930, pp. 612.

Swift, Edgar James: *The Psychology of Childhood.* New York: D. Appleton and Co., 1930, pp. X + 430.

General Psychology

Bluemel, C. S.: *Mental Aspects of Stammering.* Baltimore: The Williams and Wilkins Co., 1930, pp. VIII + 152.

DeBiran, Maine: *The Influence of Habit on the Faculty of Thinking.* Baltimore: The Williams and Wilkins Co., 1929, pp. 227.

Denison, J. H.: *The Enlargement of Personality.* New York: Charles Scribner's Sons, 1930, pp. XXII + 340.

Dodge, Raymond: *Conditions and Consequences of Human Variability.* New Haven: Yale University Press, 1931, pp. X + 162.

Gilliland, A. R., Morgan, John J. B., Stevens, S. N.: *General Psychology.* New York: Heath and Co., 1930, pp. VII + 439.

JAENSCH, E R : *Eidetic Imagery.* New York, Harcourt, Brace and Co., 1930, pp 136

RALSTON, GAGE *Present Day Psychology* Chicago J. B Lippincott Co , 1931, pp. XIV + 404

PIERCE, FREDERICK: *Dreams and Personality* New York: D Appleton and Co , 1931, pp XI + 336.

SPEARMAN, C . *Creative Mind* New York: D Appleton and Co , 1931, pp XII + 162.

WARREN, CARMICHAEL: *Elements of Human Psychology* New York: Houghton Mifflin Co , 1930, pp. VIII + 462

WATSON, JOHN B. *Behaviorism.* New York. W. W Norton and Co , 1930, pp XI + 308.

SOCIAL PSYCHOLOGY

DAVIS, SHELDON EMMOR *The Teacher's Relationships* New York. Macmillan Co., 1930, pp XIII + 415

FOLSOM, JOSEPH K . *Social Psychology* New York Harper and Bros , 1931, pp XVIII + 701

MCCASKILL, JOSEPH C. *Theory and Practice of Group Work* New York Association Press, 1930, pp VII + 165

SMITH, J J.. *Social Psychology* Boston The Gorham Press, 1930, pp XXV + 468.

WOODWORTH, ROBERT S *Contemporary Schools of Psychology.* New York: Ronald Press Co., 1931, pp VI + 231

HISTORY OF EDUCATION AND CRITIQUES

BAGLEY, WM C. *Education, Crime, and Social Progress.* New York: The Macmillan Co , 1931, pp XV + 150

BUCHHOLZ, H E *Fads and Fallacies in Present-day Education* New York The Macmillan Co , 1931, pp XIV + 200

FOSTER, RICHARD ALLEN *The School in American Literature* Baltimore Warwick and York, Inc , 1930, pp. VII + 199.

KOOS, LEONARD V *Private and Public Secondary Education* Chicago: University of Chicago Press, 1931, pp VIII + 228

PRESCOTT, DANIEL ALFRED: *Education and International Relations.* Cambridge Harvard University Press, 1930, pp. IX + 168.

REISNER, EDWARD H *The Evolution of the Common School* New York. The Macmillan Co , 1930, pp X + 590

THWING, CHARLES F.: *American Society.* New York: The Macmillan Co., 1931, pp. IX + 271.

John Dewey the Man and His Philosophy. Addresses Delivered in New York in Celebration of His Seventieth Birthday. Cambridge: Harvard University Press, 1930, pp. VII + 181

REPORTS

The Commonwealth Fund Annual Report, 1930. New York, 1931 pp. 85

FIFE, ROBERT HERNDON. *A Summary of Reports on the Modern Foreign Languages* New York: The Macmillan Co., 1931, pp. VII + 261.

SUPPLEMENTARY READERS AND CHILDREN'S BOOKS

DEARBORN, BLANCHE J.: *Kitten-kat.* New York: The Macmillan Co., 1930, pp. VI + 109

EVANS, LAWTON B.: *The Pathfinder.* New York. Macmillan Co., 1930, pp. XII + 515.

LICHTENBERGER, ANDRE. *Trott and His Little Sister.* New York. Viking Press, 1931, pp. X + 245.

LINDERMAN, FRANK B.: *American.* Yonkers-on-Hudson, N. Y. World Book Co., 1930, pp. XI + 324.

MALOT, HECTOR: *Sans Famille* Chicago University of Chicago Press, 1931, pp. XI + 134.

RENICK, DOROTHY *Star Myths from Many Lands* New York: Charles Scribner's Sons, 1931, pp. XI + 206

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII September, 1931 Number 6

OUR NEED OF SOME SCIENCE IN PLACE OF THE WORD "INTELLIGENCE"

C SPEARMAN

I CONTRIBUTIONS OF PROFESSOR DEARBORN

Not long ago the present writer ventured to criticize a book by Professor Dearborn on "Intelligence Tests"[1] He has now been good enough to respond;[2] and on one point at any rate I hasten to acknowledge a misunderstanding I had complained that his book, whilst referring copiously to the small and controversial part of a work of mine,[3] had nevertheless taken no notice of the large and constructive part. In reply, he now states that his book, although published in 1928 (June) consisted really of some old lecture notes written two years earlier, hence, he says, its contents did not refer at all to the said work of mine published in 1927 (March), but instead to another work written by me as far back as 1922.[4] Although not without surprise at his manner of making books, I must admit that it has here rendered my criticism inapplicable, and so, with apologies, I withdraw it

However Professor Dearborn, after thus making his own defence, passes over to a counter-attack I will quote his own words, in order to render justice, not only to his purport, but also to his style He charges me with "a brazen, if belated, effort to steal some of Binet's thunder," and supports this charge by alleging that "Binet did have some ideas of his own, preciously like the common factor of Spearman "

---

[1] Houghton, Mifflin and Co

[2] This Journal October, 1930

[3] "The Abilities of Man," Macmillan, 1927

[4] "Nature of Intelligence and Principles of Cognition," Macmillan, 1922

To begin with, I should like to congratulate him on his choice of objective. Certain critics of our theory of Two Factors make a pretence of attacking it fundamentally when in truth they are only dealing with unimportant details, what is really essential in our theory they tacitly appropriate to themselves (mostly under new names). Not so Professor Dearborn. The point of attack chosen by him is both novel and vital He would appear to charge us with nothing less than plagiarism; the very hub on which our theory rolls, the general factor itself, this he accredits to our earliest antagonist, Binet.

Despite this tribute to his honesty at any rate, in another respect Professor Dearborn, to speak frankly, is disappointing Since he takes up a position of such grave moment, one might expect to find him entrenched behind a great mass of specific and cogent evidence. Actually, he brings forward no evidence whatever. Although my own representation of Binet's work had from end to end been accompanied by citations of that writer's very words, Prof. Dearborn in his turn does not add a single new quotation; nor does he even indicate any fresh way of interpreting those cited by me. Instead of any such thing, he gives us simply his personal conclusions. Such a procedure may be well enough in the class room, or even in electioneering hustings But on the arena of science, just as in the courts of justice, I submit that it is too high-handed. We require the actual facts on which the assertions claim to be founded, so that we may have an opportunity to examine such claim for ourselves.

II Dilemma of Binet

Attempting to make good this default on his part, let us first of all see what was the standpoint of Binet *before* the appearance of "'General Intelligence,' objectively determined and measured,"[1] which was published by the present writer in April 1904, and already contained all the essentials of the theory of Two Factors In particular, its main object—as shown in the very title—was to introduce a "general intelligence" that admitted of measurement by tests. Before this publication, the standpoint of Binet seems to have been most characteristically expressed as follows·

> We propose the study of the ten following processes Memory, the nature of mental images, imagination, attention, the faculty of understanding, suggestibility, the aesthetic sentiment, moral sentiment, muscular force, and force of will, motor ability, and coup d'oeil. These are, we believe, mental faculties that differ

---

[1] *American Journal of Psychology*, Vol XV.

greatly from one individual to another, and are such that the knowledge of their state with an individual gives us a general idea of that individual.[1]

He then proceeds to supply us with several tests for each faculty Memory is to be tested by geometric designs, phrases, music, and colours. Mental images, by the nature of the errors of reproduction Imagination, by ink blots, composition of pictures, and combination of words into phrases Attention, by regularity of reaction-times, and by success in counting two metronomes simultaneously Comprehensions, by definitions, by perception of resemblances and differences of meaning, and by criticism of phrases.[2] Of any "general factor" he does not so far seem to have manifested any trace Indeed, even the term "intelligence" does not appear to be mentioned And to the best of my knowledge, he made no radical departure from this standpoint right up to our aforesaid publication of the theory of Two Factors in 1904.

But he certainly did so two years *after* that date. For then he and Simon brought forward their world-famous "Metric Scale of Intelligence" They wrote.

> Our aim is in no way to study, to analyse and disengage the aptitudes of those who are inferior in intelligence That will be the object of a later work Here, we will restrict ourselves to appreciating and measuring their *intelligence in general,* we will fix their intellectual level; and to give an idea of that level, we will compare it with that of normal children of the same age, or of analogous level [3]
>
> But here, we must understand one another on the meaning of that word so vague and so comprehensive: "intelligence" Almost all the phenomena with which psychology is concerned are phenomena of intelligence, a sensation, a perception, are intellectual manifestations as much as is a reasoning Ought we then to introduce into our examination the measurement of sensation, as is done by psycho-physicists? Ought we to put into our tests the whole of psychology?
>
> A little reflection has shown us that that would be a great loss of time There is *within intelligence,* so it seems to us, a *fundamental organ,* that whose lack or perversion is of the greatest importance for practical life. This is judgment, otherwise called good sense, practical sense, initiative, the faculty of adapting oneself. To judge well, to understand well, to reason well, these are the essential springs of intelligence [4]

These words, at first reading, were delightful enough. For between this measurable "intelligence in general" thus newly advocated by

---

[1] *Année Psychol*, Vol II, published in 1896 The date on the volume is 1895, but this, it appears, always refers to the *previous* year (supposedly the time that the actual research was effected)

[2] *Ibidem*

[3] *Année Psychol.*, Vol XI, published in 1906, see above.

[4] *Ibidem*,

Binet and the measurable "general intelligence" introduced two years earlier by myself, there seemed to be no difference. And as for his "fundamental organ within the intelligence," this appeared to be only a special interpretation of our factor "*g*." There are accordingly some grounds for concurring with Prof Dearborn to the extent that *by this time* (1906) Binet did have the concept of a general factor in intelligence.

But of any such likeness between his view and ours, Binet himself seems not to say a word. As I pushed on with the perusal of his work, expecting every moment to come upon his expression of cordial agreement, I encountered indeed a review of our theory of Two Factors, but this only announced his disapproval of it, on the ground of it being "stuffed with mathematics"[1]

The fact seems to be that, so far as theorizing is concerned, his adoption of our general factor was only half-hearted. He was torn in two opposite directions On the one hand, he naturally endeavoured bring his utterances into harmony with his new metric scale. But on the other hand, he continually regressed back to the old faculty doctrine, for upon this had been built up, and now stood irrevocably founded, his whole general psychological outlook Even in this very metric scale that we have been discussing, he flagrantly lapses into the facultism again; for he expressly assigns to each of his tests its own special faculty or faculties. Thus his test of cutting paper is said by him to be one of "voluntary attention, reasoning, and visual imagination" His discrimination between objects is called a test of "ideation, the notion of difference, and to some extent the spirit of observation" His repetition of digits is said to involve the faculties of "immediate memory and of voluntary attention" His test of pictures—to which he subsequently attributed such dominant importance—is here only taken as "destined to seek whether there exist associations between images and their names" In the second version of the metric scale published three years later, all such indications of special faculties to be measured by the respective tests was, indeed, abandoned. But very soon afterwards, they were re-introduced and more explicitly than ever. For now he goes so far as to write that:

> The mental faculties of each subject are independent and unequal Our mental tests are always specific in their scope; each suits the analysis of a single faculty.[1]

---

[1] Les Idées modernes sur les Enfants, 1909

In these words, the general factor or "fundamental organ" would seem to have quite faded out of the picture.

But whilst thus in theoretical discussion ever regressing back to his ingrained facultism in the actual practice of measuring the logic of facts drove him irresistibly more and more along the new path of Two Factors Already in 1900, as mentioned, he had indicated his intention to measure the "level" of his "intelligence in general" But how was such a feat possible? The word "general" means something common to all or nearly all of a class, as exemplified in the phrases a "general direction" or the "general government." But with the faculties, he himself stated (as we have just seen) that no such common level exists His sole resource was to take "general" not in the sense of common, but in that of middle or average Such averaging, however, was just what we had advocated already in 1904. Brilliantly original, then, as was his procedure in detail, in essence it was following us

III. Limitations to the Validity of Averaging

But was he here really entitled to follow us? The procedure of averaging, wide-spread and invaluable as it is in both physical and mental science, has nevertheless its limitations It is only legitimate when certain postulates are fulfilled One of these is the unequivocality of the domain at issue You cannot, for instance, obtain the average height of an American without first settling who is to count as one And then, having got the domain definitely settled, it is imperative to take the whole of this into the reckoning. Thus, you cannot claim to have got the monthly average of your expenditure during any year if you leave out of calculation, say, December A third postulate is that your reckoning should present no overlaps, thus, it would not do to bring in December several times Still more fundamental is the need of comparable units. You may indeed equally well derive your average expenditure by taking every day, or every week, or every month. But you would not be rational if you mixed up days, weeks, and months indiscriminately together.

Now, on the theory of faculties, one at least of these postulates appears impossible to satisfy, namely, that of finding any common unit by which the different faculties can be brought into mutual connection Thus, how shall a common unit be found to make memory comparable with judgment? We might take the whole of memory as one single unit whilst we divided up judgment into several kinds each constituting a unit of its own; or else equally well we might take judgment as being one and subdivide the memory. The one basis for averaging would favor some individuals, and the other basis others Thus the score reached by averaging becomes arbitrary, therefore invalid

And the matter becomes much worse still, if we pass from the clear but untenable theory of faculties to the extremely vague theory of "intelligence" commonly favored. For here, all four of the said postulates fail to be satisfied To find any common unit remains as impossible as before. In addition, there seems no way of avoiding large omissions; for it would appear hopeless to attempt to include in the tests every sort and description of "intelligence." Furthermore, there appear to be no means of arranging that none of the tests overlap each other, indeed, they often admittedly do so And lastly, no agreement seems to be feasible as to how the domain of this "intelligence" should be delimited, for some authorities want to include one thing and others others.

Nor would the preceding difficulties be in any degree escaped by shifting from the notion of average to that of a "totality" For the latter also would involve every one of the aforesaid postulates (even—as might perhaps not be expected—that of a common unit). Nor yet would any relief be brought by substituting, as many later writers have done, the notion of a "sample." For this too would involve the same four postulates—with other troubles superposed.[1] To all this may be added that actually mental testers, far from having overcome these difficulties, have hitherto not so much as made any serious attempt at doing so.

But what is sauce for the goose, it may be urged, is also sauce for the gander. If all this averaging is so irrational when employed by the facultists and the intelligentists, how can it pretend to be satisfactory in the hands of the two-factorists? The reply is that in the latter case all the preceding difficulties do surprisingly dissolve away. For in it, no claim is raised that the result of combining together different tests is to constitute any rational average, or central value, or totality, or sample The sole claim raised is that, in such a procedure, the specific factors tend to cancel out, leaving the general factor alone dominant As was said originally:

> "The specific element can to a great extent be readily eliminated by varying and combining the kinds of test," whereby the general factor or "central function" stands out in corresponding exactness [2]

---

[1] "The Abilities of Man," Chap. V

[2] *American Journal Psychology*, Vol XV, April, 1904

For *this* purpose of cancelling out, omissions or overlappings—unless they are highly systematic—do not greatly interfere Nor does even the manner in which the units are constituted. And as for the domain from which the tests are to be derived, this in the theory of two factors is no longer equivocal. Its domain extends to just as many abilities as may be observed to satisfy the two-factor criterion.

IV. PLEA FOR AN AWAKENING

What is the upshot of all this? Not unnaturally perhaps, Prof Dearborn comes to the conclusion that, in spite of all, the situation is well enough; that, even if the current procedure of measuring the "mental ages" and the "IQ's" is indeed equipped with a wrong theory, it nevertheless admittedly is right in practice, which is what alone really matters

Now, it is just against this attitude of quietism that I would here desire to make a lively protest For in the first place, even the practice current in testing is only right with very crude approximation. It only succeeds ever in yielding a definite measurement at all if and when the sub-tests entering into the whole team of them are sufficiently numerous and *selected at random* Now some sort of randomness does derive from the mere fact that these subtests are gathered together from highways and byways without any system or plan But this facile mix-up is insufficient to secure that perfect randomness—in the sense of freedom from all bias—which statisticians mean by the term and which is needed to produce the definite "*g*" In order to obtain such freedom from bias, the procedure should not be without all plan, but on the contrary planned with great care Whereas the present test-makers, quiet light-heartedly allow themselves to be strongly biassed towards several objectives, such as ease of application, self-consistency, correlations with other things, and so forth.

But there is a still worse objection to the current procedure Even if this did happen to produce pure *g* (or any other definite measure), nevertheless when divorced from the theory out of which it sprang it would pay the penalty of forfeiting all significance; the mere average of sub-tests picked up and put together without rhyme or reason must needs present—at any rate until proof is given otherwise—the very acme of *meaninglessness*. And as for the customary dignifying of this gallimaufrey of tests by the title of "intelligence"—without so much as an attempt really to construct them in accordance with any explicit principle—this device comes unpleasantly near to psychological characteristics.

Having regard to these two defects, then, those of indefiniteness and of meaninglessness, there is but small wonder that the current procedure from time to time commits, even in practice, blunders of great magnitude, so that there is only too good reason to suspect that many a child branded by test has thereby unjustly had his career wrecked And this terrible fallibility of the tests is only rendered the more dangerous by the vast edifice of statistical data and calculations which are often ostentatiously piled upon them and which do but convert them into whited sepulchres

But by what manner of means, it may be asked, is all this evil going to be made any the better by bringing on the scene the theory of Two Factors? Are not these also repeatedly charged with being entities of an obscure, speculative, or hypothetical kind? To say this about the factors is totally to misapprehend them, as we will now endeavour to show. These two factors, g and s, do not in themselves signify any entities whatever; they are nothing more than parts of actually observed measurements of ability And these part-values have at least the virtue of perfect quantitative definiteness, that is to say they are, when taken rightly, "unique" But to acquire scientific *meaning, a further step is required, there must be proof supplied that* these part-values, besides being definite quantitatively in themselves, also serve as measures of some qualitatively definite psychological functions And this, too, we may claim, has actually been done. In particular, the g has been found to measure "noegenesis"[1] This term denotes the following three actually and precisely observable mental functions The first is the knowing of our own experience: we not only feel, but know that we feel, we not only strive, but know that we strive, we not only know, but know that we know. The second of the mental functions consists in that, when any two or more items are perceived or thought of, we may cognize relation between them, for example, we may become aware that one thing is like another, or is evidence for another, or is the cause of it An outstanding instance of this among tests is that of "synonyms and antonyms," where two words are given and the testee has to decide whether they are very alike or very unlike. Thirdly and lastly, when we perceive or think of any item and also of any appropriate relation, we may evoke the idea of the item which stands to the given one in the given relation. A well know instance among tests is that of "opposites," where the testee is given some word and is told to say what is the opposite to it These second and third functions are usually characterised as "eductive"

---

[1] The "noe-" comes from the Greek νοῦς, and indicates that processes of this form alone produce knowledge The "-genesis" indicate that such processes alone generate new cognitive content

Whilst thus urging that the noegenetic processes are conceived in a manner free from all indefiniteness, we may take the opportunity to defend them from the charge sometimes raised that they are only of a "logical" nature, not of a "psychological" one What is here meant by logical has been left by these critics regrettably obscure But possibly the notion is that these kinds of process have little or no importance for understanding, anticipating, and controlling the course of mental life. In reply, we must claim just the contrary These three processes follow in general a settled sequence, and in such fashion that each prepares the way for the following one. Thus, the awareness of one's own experience supplies all the fundamental items of knowledge; next, between these items (possibly in some cases simultaneously with them) the relations are cognised, and finally the relations thus furnished are displaced to other items and thus produce all correlative and novel items which are the mainstay of mental creation and invention, including that super-invention which is behavior.[1] To recognise this sequence of the three steps and to explore the conditions which facilitate each of them should constitute one of the main objectives of all functional psychology, including in particular its applications to education, medicine, and industry. And all this is called "not psychological but logical"!

More subtle and therefore still more dangerous than any such direct charges against the noegenetic analysis is the not uncommon practice of tacitly ranking it with mere unevidenced assertions on the matter For example, the view that mental tests involve noegenesis is put on the same footing with the assertion that they involve adaptibility or educability, and so forth This is done regardless of the fact that the noegenetic view is based on the most elaborate and exhaustive measurements in all detail, whereas the other views have nothing behind them but unsupported statement

But even this establishment of g as being a measure of the ability called noegenesis is only the beginning, by no means the sum, of the definite and scientifically significant contributions of the theory of Two Factors. In support of this claim, appeal may be made to the second part of "The Abilities of Man" quoted on the first page of this paper. For this whole part is packed, not with anything obscure or speculative, but with fundamental facts actually observed by a multitude of indefatigable workers during the past quarter of a century And to all this light which the Two Factors have shed upon the working of the mind must be added that which they have thrown upon the current tests of the socalled general intelligence For the *g* appears to constitute just that which is common to different tests or sets of tests, and thus to furnish the sole stable element in intelligence-testing. To this stable element, the tests at present in use supply only crude approximations; what they contain over and above this *g* varies in an incalculable and therefore unscientific manner from one such test to another

---

[1] For a full but simple account of such displacement of relations, reference must be made to the present writer's "Creative Imagination," Appleton, 1930

There is yet another objection commonly raised to the theory of two factors Here, the claim is advanced that this mode of analysing the scores made in a test is only permissive, not obligatory; other manners of analysis are proposed—or more often vaguely suggested as possible—to take its place. Against this attitude we must plead that, on the contrary, there are places for any number of different analyses; any two different ones may quite well be *both of them* true and even both of them valuable, or both of them the reverse. Each has to stand or fall independently, its fate must be settled solely by the amount of scientific information which it produces For example, the definitely observed fact that *g* measures noegenesis remains quite unaffected by what any other factors arising from any different analysis may or may not happen to measure. If any factors other than *g* and *s* do also happen to render useful services, then these services can only supplement one another, and thus enhance each other's values

All in all, then, the plea is here raised that the current procedure of testing "intelligence" needs to be aroused from its self-complacent slumber. The present real basis of it is only a practice that has been borrowed from another theory. And this practice fails to be satisfactory because, when the practice was borrowed, the theory itself was left behind. By so doing, a sacrifice was made of precision; of meaningfulness, and above all, of an immense amount of observed fundafental facts This disastrous situation has been largely masked, but at the same time really aggravated, by usurping the pretentious and wellnigh fraudulent title of "intelligence."

AN EXPERIMENTAL STUDY OF THE INFLUENCE OF MOTION PICTURE FILMS ON BEHAVIOR[1]

FRANK N FREEMAN

University of Chicago

AND

CAROLYN HOEFER

Elizabeth McCormick Memorial Fund

The purpose of this study was to determine whether motion picture films add to the effectiveness of instruction which is designed to incite children to perform specific acts. The problem is one phase of the more general problem of the comparative effect of motion pictures and of other influences in guiding or modifying conduct. In this particular instance the instruction, of which the motion pictures are a part, was designed to influence conduct largely through giving clear-cut information and pointing out the applications of this information to conduct The motion pictures which were used should have been chiefly of an informational nature to harmonize with the character of the general unit of instruction Suitable informational films could not be obtained, however, and the films which were used were cast in story form. They were designed to appeal to the emotions as well as to the intellect

The instruction dealt with the care of the teeth. This subject was chosen because it was believed that any change in the children's behavior as a result of instruction could be more objectively measured in this field than in any other which was open to investigation The Elizabeth McCormick Memorial Fund, moreover, was in a position to provide systematic curriculum material on this subject and to make contacts with officials of the Health Department of Chicago and of the Public Schools as a result of the co-operative study of other problems in health education.

---

[1] This study was made possible by a grant from the Committee for the Study of the Social Values of Motion Pictures of the Payne Fund Grateful acknowledgment is also made to Dr George Wandel of the American Dental Association and Dr. Lon Morrey of the Health Department of Chicago for the loan of the films and for assistance in the preparation of dental lessons, to the Chicago Department of Health and the Superintendent of Schools for the assistance given in the dental examinations

The experiment was made in two public schools of Chicago Two fifth and two sixth grades in one school were taught with the aid of the films, while two fifth and two sixth grades in another school in a comparable neighborhood were taught without the films In order that the instruction might be comparable, except for the films, it was all given by the same person, Miss Mildred Dawson. Miss Dawson also served as general assistant in the experiment. She is a very well-trained and competent teacher and supervisor.

The instruction was given in thirteen periods, from May 8 to May 22 inclusive. The following is the outline of topics·

Lesson 1 —Know your teeth.
Shape and function of teeth
Lesson 2 —Structure of a tooth.
Lesson 3 —Diet for teeth.
Lesson 4 —Diet for teeth.
Lesson 5 —Deciduous teeth.
Lesson 6 —Permanent teeth.
Lesson 7 —Six-year molars
Lesson 8 —Film or story.
Lesson 9 —Malocclusion and mastication.
Lesson 10.—Causes of decay
Lesson 11 —Prevention of decay, and review
Lesson 12 —Care of the teeth, and review
Lesson 13 —Film or writing slogans

Charts and diagrams were used in the instruction of both groups to supplement oral instruction. The time occupied by showing the film to the experimental group was employed in reading a story and in conducting an exercise in writing slogans in the control group

The two films which were shown to the experimental group are entitled Tommy Tucker's Tooth and Clara Cleans Her Teeth The first tells the story of a boy who was refused a job for which he applied because of his unsightly teeth, but who later had his teeth looked after and was then accepted The second tells of a girl who enters a new school where the children have been trained to care for their teeth, and who is ostracized because of the poor condition of her teeth and her indifference to it. She, too, reforms and becomes socially acceptable to her schoolmates. Some information is given in the films, but it is subordinate to the stories

The effectiveness of the instruction in the experimental and control groups was measured in four ways, namely:

1 An information test, given at the beginning and the end of the period of instruction,

2 A questionnaire in which the children reported on the care of their teeth at home and the food which they ate,

3 The material brought in by the children, consisting of clippings of various sorts, original material written by the children, booklets read, literature taken home, and score cards which the children filled out;

4 Condition of the teeth as determined by a dental examination given at the beginning and at the end of the experiment

RESULTS THE INFORMATION TEST

The information test was given at the beginning and the end of the experiment to both groups It consists of thirty-six multiple-choice questions. The questions relate to the structure and the development of the teeth, the causes of defective teeth, and the care of the teeth Through difficulty in tabulation, four of the questions were omitted, leaving thirty-two The information test is not a direct measure of behavior Since the purpose of much of the instruction, however, was to give information, it is important to know whether the films had any influence directly or indirectly upon the amount of information which the children obtained A comparison of the gains in information in the two groups is, therefore, presented

The differences in the average ability of the children in the individual grades of the two schools was so great that it was thought necessary to make matched groups based on scores on the intelligence tests and the initial scores on the information test The matched groups were formed, not by matching individuals, but by selecting individuals from the two groups so that there would be an approximately equal number in each class of the distribution This form of matching gives not only equal average scores, but also an approximately equal number at the various levels of ability

Table I gives the data on the results of the information test from the equalized groups It will be observed that the average intelligence quotient and the average initial score of the two groups are practically identical

It is apparent that both the boys and girls of the control group made somewhat higher average scores on the information test than did the children of the experimental group, but the difference is hardly statistically significant. We are safe in saying, however, that the films had no appreciable influence on the acquisition of information This is, perhaps, not surprising, since the films were not designed primarily to give information. They were designed rather to make the children interested in caring for their teeth. All we can say, then, is that these films, designed to awaken interest in the care of the teeth, did not stimulate the children who saw them to acquire more information than the children who did not see them. Interest in acquiring information about the teeth is evidently somewhat independent of interest in the care of the teeth, assuming that the films have stimulated interest in the care of the teeth. It is, of course, possible that they did not stimulate such interest. The inference that the two types of interest are independent, therefore, is not conclusive. It is contingent upon an assumption for which we have as yet no evidence. We shall have to hold the conclusion in abeyance, therefore, until we have examined the evidence concerning the effect of the films in stimulating interest in the care of the teeth.

TABLE I.—THE SCORES OF THE EQUALIZED EXPERIMENTAL AND CONTROL GROUPS ON THE INFORMATION TEST AT THE BEGINNING AND THE END OF THE EXPERIMENT

Equalized groups[1]	No.	Average IQ	SD	Average Test I	SD	Average Test II	SD
Experimental school							
Boys	55	100.98	15.247	14.91	3.561	27.73	5.992
Girls	53	101.32	11.474	15.40	4.003	28.62	4.180
Boys and girls	108	101.15	13.517	15.15	3.793	28.17	5.103
Control school.							
Boys	50	100.78	16.364	14.90	3.689	29.64	5.768
Girls	49	101.37	11.642	15.41	3.523	29.06	4.895
Boys and girls	99	101.07	14.235	15.15	3.022	29.35	5.370

[1] Equalized on basis of sex, intelligence rating and initial score on information test.

In order to discover whether the film was of special aid in giving particular items of information, the individual questions of the information test, which bore upon the activities involved in the care of the teeth, were tabulated separately. In only two questions does the experimental group give any clear evidence of being superior to the control group in the gains on the individual questions. The correct answer to the first of these questions is: The motion of the brush should be up and down, and to the second: To keep the teeth in good condition we should eat some food that requires much chewing. The method of using the toothbrush was shown in both films, and the use of food which requires chewing was shown in the film, Tommy Tuckers' Tooth The superiority of the film group in these two questions, therefore, is plausibly explained by the fact that the items of information or of activity to which they refer are specifically emphasized in the film

THE RESULTS OF THE QUESTIONNAIRE

At the end of the experiment, the children of both groups were given a score card to fill out On this score card they were asked to report whether or not (1) they possessed a tooth-brush, (2) they brushed their teeth daily, (3) they used a dentifrice, (4) they ate vegetables daily, (5) they ate fruit daily, and (6) they drank milk daily.

The results of the questionnaire are tabulated in Table II. The table shows the percentage of children of each group who answer the various questions in the affirmative. It is obvious that, if the children's testimony is to be relied upon, a large percentage of both groups perform faithfully the act which they were taught to perform in the instruction Whether or not they would have given an equally good account of themselves at the beginning of the experiment we do not know. On account of the high scores in both groups and the uncertainty concerning the accuracy of the children's replies, small differ-

TABLE II.—THE RESPONSES TO THE QUESTIONNAIRE ON THE CARE OF THE TEETH AND ON EATING GIVEN AT THE END OF THE EXPERIMENT

		Home care						Food					
	No	Possession of tooth-brush		Teeth brushed daily		Use of dentifrice		Vegetables eaten daily		Fruit eaten daily		Milk drunk daily	
		No	Per cent	No.	Per cent	No.	Per cent	No.	Per cent	No.	Per cent	No	Per cent
Experimental school[1]													
Boys	92	92	100 00	90	97 83	88	95 65	85	92 39	88	95 65	91	98 91
Girls	100	99	99 00	98	98 00	95	95 00	94	94 00	91	91 00	95	95 00
Boys and girls	192	191	99 48	188	97 92	183	95 31	179	93 23	179	93 23	186	96 88
Control school.[2]													
Boys	89	86	96.63	66	74 16	83	93 26	81	91 01	87	97 75	89	100 00
Girls	81	80	98.77	72	88 89	77	95 06	74	91 36	78	96 30	77	95 06
Boys and girls	170	166	97 65	138	81 18	160	94 12	155	91 18	165	97 06	166	97 65

[1] Records of ten boys and six girls incomplete due to absence

[2] Records of eight boys and four girls incomplete due to absence

ences between the two groups are of no significance. The only item on which the difference may be significant is that concerning the daily brushing of teeth. Ninety-eight per cent of the children of the experimental group reported that they brushed their teeth daily whereas only eighty-one per cent of the control group made a similar report. Since the films emphasize the daily brushing of teeth very strongly, this difference may be significant The films also emphasize the other items, however, and it is not altogether clear why there should be so much more difference in this item than in the others. So far as it goes, the data suggest that when a particular action is strongly emphasized in a film the emphasis may be effective However, it is not at all clear why the control group should have fallen so much lower in brushing their teeth daily than in using dentifrice, or in eating the various kinds of food which were advised Taken altogether, then, the results of the questionnaire do not indicate much one way or the other concerning the relative influence of the motion picture and of oral instruction.

Measures of Interest

As measures of interest a record was kept of certain specific voluntary activities on the part of the children This record included, first, a record of all the clippings or original articles, pictures, poems, stories, etc., which were contributed by the children, second, the number of booklets which were read by the children of their own accord, the amount of literature which was taken home and the number of score cards concerning their reading which they voluntarily kept. The booklets, literature, and score cards were placed on the desk where the children could get them and they were told that they were free to use them if they wished to do so.

A summary tabulation of the contributions and voluntary acts is found in Table III. A comparison of the integral numbers in the table is qualified somewhat by the difference in the total number of pupils in the two groups. The number represented in the experimental group is 208 and in the control group 182. Some of the differences, however, are of such magnitude that this qualification is of minor importance

It will be seen that the pupils of the experimental group brought in a much larger number of articles and pictures and contributed a larger number of original articles, cartoons, and pictures The pupils of the control group brought in a somewhat larger number of poems and drew a somewhat larger number of posters The differences, however, are not so large as those in favor of the experimental group. Taken as a whole, it is clear that the pupils of the experimental group

Table III.—A Record of Certain Voluntary Activities Carried on, Used as a Measure of Interest

	Experimental school (102 boys, 106 girls, 208 total) No.	Control school (97 boys, 85 girls, 182 total) No
Contributions		
Clippings.		
Article	53	7
Cartoon	3	5
Chart	7	1
Magazine	1	3
Model	0	1
Pamphlet	2	0
Picture	410	72
Poem	2	29
Poster	0	4
Story	2	1
Total	480	123
Average per child	2.31	0.68
Original		
Article	16	3
Cartoon	25	4
Chart	1	4
Picture	40	9
Poem	10	4
Poster	0	14
Story	7	1
Total	99	39
Average per child	0.48	0.21
Grand Total	579	162
Average per child	2.78	0.89
Average booklets read	0.69	0.66
Average pieces of literature taken home	0.74	0.82
Average score cards filled out	0.46	0.35

contributed a much larger number of articles and of original materials of various sorts as reckoned in terms of the average number of contributions per child In the number of booklets read, the literature taken home, and the score cards filled out, however, there is no significant difference between the two groups. The balance is slightly in favor of the experimental group and the influence of the film appears to be somewhat significant It is probably significant that the item in which the experimental group is most superior is in the number of pictures which were contributed. The sight of the motion picture apparently suggested to the children that they find similar pictures and bring them to the class The results do not suggest a great difference in general interest or motivation between the two groups.

CHANGE IN THE CONDITION OF THE TEETH

It was thought that a change in the condition of the teeth might be a more objective measure of the care which was exercised by the children or their parents and might, therefore, be the most objective measure of the effectiveness of instruction The two types of changes which seem to be most significant are the change in the number of carious teeth and the change in the number of teeth which have received dental care The reports of the dental examinations at the beginning and the end of the period of instruction were therefore examined to obtain a comparative measure of these two types of changes in the experimental and the control groups.

It is necessary to confine the comparison to the permanent teeth There is a difference in practice among dentists in the treatment of deciduous teeth—some fill them, others do not.

Again, some deciduous teeth disappear between the first and second examination and the number of carious teeth would, therefore, be decreased without any effort on the part of the child or his parents.

The first permanent molars, sometimes referred to as six-year molars, were selected for examination because these are the permanent teeth which have been in the mouth the longest time They, therefore, give the greatest opportunity for decay and for the exercise of dental care.

In interpreting the condition of the teeth or the change in the condition of the teeth it is necessary to take careful account of age The eruption of these molars themselves is highly correlated with age and the eruption of the neighboring teeth, which have an influence upon decay, is also correlated with age For this reason it was necessary to equalize the groups in accordance with the distribution of ages, in the same manner as was done for intelligence and initial score in the tabulation of the results of the information test Since the teeth of girls erupt earlier on the average than those of boys, it is necessary also to keep the results from the two sexes separate in the tabulation

The first tabulation of a change in the condition of the teeth is given in Table IV. This table shows the increase or decrease in the number of carious teeth from the first to the second examination. The groups have been equalized with respect to chronological age, but obviously they have not been equalized with respect to the condition of the teeth at the first examination In the experimental group there were 253 carious molars at the first examination whereas in the control group there were 318 This may indicate greater susceptibility of the children of the control group to caries or it may indicate that these children have had less care The average number of carious molars per child increased in the experimental group and decreased in the control group. At first sight this may appear to indicate that the instruction was more efficient in the control group than in the experimental group. Account must be taken, however, of the difference in the condition of the teeth of the two groups at the beginning of the experiment It is at least possible that the greater increase in the carious teeth of the children in the experimental school is due to the fact that there was greater opportunity for increase

TABLE IV —CHANGE IN THE NUMBER OF CARIOUS FIRST PERMANENT MOLARS FROM THE FIRST TO THE SECOND EXAMINATION AFTER THE GROUPS HAVE BEEN EQUALIZED FOR AGE

	No.	Average CA months	First examination		Second examination	
			No. carious *molars*	Average per *child*	No carious *molars*	Average per *child*
Experimental school.						
Boys. . .	80	138 37	122	1 42	137	1 59
Girls	81	135 64	121	1 49	126	1 56
Boys and girls	167	137 04	243	1 46	263	1 57
Control school						
Boys	80	137 70	168	1 95	167	1 94
Girls	81	135 56	150	1.85	133	1 64
Boys and girls	167	136 68	318	1 90	300	1 80

In order to furnish a check on this possibility, the groups were equalized not only with respect to chronological age, but also with respect to the condition of the first permanent molars at the first examination. In other words, the children were selected in the two groups so as to provide an equal number of carious molars at first examination. The comparison is shown in Table V. The difference in the two groups is now reversed. The number of carious teeth per child decreases in both groups and slightly more for the experimental than for the control group. Apparently the unfavorable showing of the experimental group in comparison to the control in the earlier table was due to a difference in the initial condition of the teeth. The difference in favor of the experimental group in the equalized groups is hardly great enough to have statistical significance, particularly as it is all due to a difference in the case of the boys. It may be that we should expect no difference. Even the best of care could probably affect but little change in the period of the experiment and the difference due to a difference in the effectiveness of the instruction in the two groups might be so minute as to be imperceptible when

TABLE V.—CHANGE IN THE NUMBER OF CARIOUS FIRST PERMANENT MOLARS FROM THE FIRST TO THE SECOND EXAMINATION AFTER THE GROUPS HAVE BEEN EQUALIZED FOR INITIAL CONDITION OF THESE MOLARS

			First examination		Second examination	
	No.	Average CA	No. of first permanent molars carious	No. carious per child	No. of first permanent molars carious	No. carious per child
Experimental school						
Boys	47	139.02	92	1.96	79	1.68
Girls	45	133.20	95	2.11	86	1.91
Boys and girls	92	136.17	187	2.03	165[1]	1.79
Control school						
Boys	47	138.94	91	1.94	85	1.81
Girls	45	133.82	95	2.11	86	1.91
Boys and girls	92	136.43	186	2.02	171[2]	1.87

[1] Twenty-two molars carious at first examination and in "perfect condition" at second examination.

[2] Fifteen molars carious at first examination and in "perfect condition" at second examination.

TABLE VI.—NUMBER OF FIRST PERMANENT MOLARS WHICH WERE IN "PERFECT CONDITION" AT THE FIRST EXAMINATION AND WHICH WERE CARIOUS, FILLED OR EXTRACTED AT THE SECOND EXAMINATION

	No	Experimental school											Control school										
		No	Av CA	Av IQ	Carious		Filled		Extracted		Received dental care		No	Av CA	Av IQ	Carious		Filled		Extracted		Received dental care	
					No	Per cent	No	Per cent	No	Per cent	No	Per cent				No	Per cent	No	Per cent	No	Per cent	No	Per cent
No. 3 Not Carious First Examination																							
Boys	86	27	137.63	101.00	6	22.2	1	3.7	1	3.7	2	25.0	19	138.47	101.11	3	15.8	1	5.3	1	5.3	2	40.0
Girls	81	28	136.04	103.93	5	17.9	4	14.3	0	0.0	4	44.4	19	137.53	96.42	6	31.6	1	5.3	0	0.0	1	14.3
Boys and girls	167	55	136.82	102.49	11	20.0	5	9.1	1	1.8	6	35.3	38	138.00	98.76	9	23.7	2	5.3	1	2.6	3	25.0
No. 14 Not Carious First Examination																							
Boys	86	27	138.00	101.07	5	18.5	1	3.7	0	0.0	1	16.7	16	135.19	103.38	3	18.8	0	0.0	0	0.0	0	0.0
Girls	81	22	135.36	104.95	4	18.2	1	4.5	0	0.0	1	20.0	21	134.24	100.95	5	23.8	0	0.0	0	0.0	0	0.0
Boys and girls	167	49	136.82	102.82	9	18.4	2	4.1	0	0.0	2	18.2	37	134.65	102.00	8	21.6	0	0.0	0	0.0	0	0.0
No. 19 Not Carious First Examination																							
Boys	86	20	142.25	96.85	4	20.0	0	0.0	1	5.0	1	20.0	18	138.50	98.22	8	44.4	1	5.6	0	0.0	1	11.1
Girls	81	17	132.35	103.47	4	23.5	0	0.0	2	11.8	2	33.3	10	134.30	98.00	2	20.0	1	10.0	0	0.0	1	33.3
Boys and girls	167	37	137.70	99.89	8	21.6	0	0.0	3	8.1	3	27.3	28	137.00	98.14	10	35.7	2	7.1	0	0.0	2	16.7
No. 30 Not Carious First Examination																							
Boys	86	21	139.05	101.14	6	28.6	1	4.8	0	0.0	1	14.3	20	137.55	103.00	7	35.0	0	0.0	1	5.0	1	12.5
Girls	81	13	136.92	101.23	2	15.4	1	7.7	1	7.7	2	50.0	13	134.62	103.85	4	30.8	3	23.1	0	0.0	3	42.9
Boys and girls	167	34	138.24	101.18	8	23.5	2	5.9	1	2.9	3	27.3	33	136.39	103.33	11	33.3	3	9.1	1	3.0	4	26.7

In order to furnish a check on this possibility, the groups were equalized not only with respect to chronological age, but also with respect to the condition of the first permanent molars at the first examination. In other words, the children were selected in the two groups so as to provide an equal number of carious molars at first examination. The comparison is shown in Table V. The difference in the two groups is now reversed. The number of carious teeth per child decreases in both groups and slightly more for the experimental than for the control group Apparently the unfavorable showing of the experimental group in comparison to the control in the earlier table was due to a difference in the initial condition of the teeth The difference in favor of the experimental group in the equalized groups is hardly great enough to have statistical significance, particularly as it is all due to a difference in the case of the boys It may be that we should expect no difference Even the best of care could probably affect but little change in the period of the experiment and the difference due to a difference in the effectiveness of the instruction in the two groups might be so minute as to be imperceptible when

TABLE V—CHANGE IN THE NUMBER OF CARIOUS FIRST PERMANENT MOLARS FROM THE FIRST TO THE SECOND EXAMINATION AFTER THE GROUPS HAVE BEEN EQUALIZED FOR INITIAL CONDITION OF THESE MOLARS

	No	Average CA	First examination		Second examination	
			No. of first permanent molars carious	No carious per child	No of first permanent molars carious	No carious per child
Experimental school.						
Boys	47	139.02	92	1 96	79	1 68
Girls	45	133 20	95	2 11	86	1 91
Boys and girls.	92	136 17	187	2 03	165[1]	1 79
Control school.						
Boys	47	138.94	91	1 94	85	1 81
Girls	45	133 82	95	2 11	86	1 91
Boys and girls.	92	136 43	186	2 02	171[2]	1 87

[1] Twenty-two molars carious at first examination and in "perfect condition" at second examination.

[2] Fifteen molars carious at first examination and in "perfect condition" at second examination

TABLE VI —NUMBER OF FIRST PERMANENT MOLARS WHICH WERE IN "PERFECT CONDITION" AT THE FIRST EXAMINATION AND WHICH WERE CARIOUS FILLED OR EXTRACTED AT THE SECOND EXAMINATION

	No	Experimental school											Control school										
		No	Av CA	Av IQ	Carious		Filled		Ex-tracted		Received dental care		No	Av CA	Av IQ	Carious		Filled		Ex-tracted		Received dental care	
					No	Per cent	No	Per cent	No	Per cent	No	Per cent				No	Per cent	No	Per cent	No	Per cent	No	Per cent
No 3 Not Carious First Examination																							
Boys	86	27	137 63	101 00	6	22 2	1	3 7	1	3 7	2	25 0	19	138 47	101 11	3	15 8	1	5 3	1	5 3	2	40 0
Girls	81	28	136 04	103 93	5	17 9	4	14 3	0	0 0	4	44 4	19	137 53	96 42	6	31 6	1	5 3	0	0 0	1	14 3
Boys and girls	167	55	136 82	102 49	11	20 0	5	9 1	1	1 8	6	35 3	38	138 00	98 76	9	23 7	2	5 3	1	2 6	3	25 0
No 14 Not Carious First Examination																							
Boys	86	27	138 00	101 07	5	18 5	1	3 7	0	0 0	1	16 7	16	135 19	103 38	3	18 8	0	0 0	0	0 0	0	0 0
Girls	81	22	135 36	104 95	4	18 2	1	4 5	0	0 0	1	20 0	21	134 24	100 95	5	23 8	0	0 0	0	0 0	0	0 0
Boys and girls	167	49	136 82	102 82	9	18 4	2	4 1	0	0 0	2	18 2	37	134 65	102 00	8	21 6	0	0 0	0	0 0	0	0 0
No 19 Not Carious First Examination																							
Boys	86	20	142 25	96 85	4	20 0	0	0 0	1	5 0	1	20 0	18	138 50	98 22	8	44 4	1	5 6	0	0 0	1	11 1
Girls	81	17	132 35	103 47	4	23 5	0	0 0	2	11 8	2	33 3	10	134 30	98 00	2	20 0	1	10 0	0	0 0	1	33 3
Boys and girls	167	37	137 70	99 89	8	21 6	0	0 0	3	8 1	3	27 3	28	137 00	98 14	10	35 7	2	7 1	0	0 0	2	16 7
No 30 Not Carious First Examination																							
Boys	86	21	139 05	101 14	6	28 6	1	4 8	0	0 0	1	14 3	20	137 55	103 00	7	35 0	0	0 0	1	5 0	1	12 5
Girls	81	13	136 92	101 23	2	15 4	1	7 7	1	7 7	2	50 0	13	134 62	103 85	4	30 8	3	23 1	0	0 0	3	42 9
Boys and girls	167	34	138 24	101 18	8	23 5	2	5 9	1	2 9	3	27 3	33	136 39	103 33	11	33 3	3	9 1	1	3 0	4	26 7

measured by the rather slow process of dental decay. Substantial equality of the two groups in this respect, therefore, is perhaps just what should be expected.

A more delicate measure of the effect of instruction is the amount of dental care which was given the teeth which needed such care This was measured by recording the number of teeth which were in perfect condition at the beginning of the experiment and became carious during the experiment These teeth evidently presented new problems in dental care. To determine how effectively these problems were met, a record was further kept of the number of such teeth which were filled or extracted during the period of the experiment This record is tabulated for each of the four six-year molars separately This tabulation is presented in Table VI. The individuals are selected from the groups which were equalized by age and approximately by average IQ. It may be read as follows: Among the eighty-six boys who made up the equalized group in the experimental group there were twenty-seven perfect teeth on the first examination On the second examination it was found that there were six carious teeth, one filled and one

TABLE VII.—SUMMARY OF NUMBER OF FIRST PERMANENT MOLARS BECOMING CARIOUS DURING THE EXPERIMENTAL PERIOD WHICH WERE GIVEN DENTAL CARE

	Summary of first permanent molars			Second examination					
	Not carious first examination	Carious		Filled		Ex-tracted		Received dental care	
		No	Per cent	No	Per cent	No	Per cent	No.	Per cent
Experimental school									
Boys	95	21	22.1	3	3.2	2	2.1	5	19.2
Girls	80	15	18.8	6	7.5	3	3.8	9	37.5
Boys and girls	175	36	20.6	9	5.1	5	2.9	14	28.0
Control school									
Boys	73	21	28.8	2	2.7	2	2.7	4	16.0
Girls	63	17	27.0	5	7.9	0	0.0	5	22.7
Boys and girls	136	38	27.9	7	5.1	2	1.5	9	19.2

extracted. In other words, eight teeth had become carious and of these two had received dental care. Twenty-five per cent of the teeth which had become carious, therefore, had received dental care The significant figures, therefore, are the percentages under the column, Received Dental Care.

The main facts of Table VI are summarized in Table VII It appears, for example, that among the boys of the experimental group there were ninety-five six-year molars which were not carious at the first examination Of these, twenty-six became carious, and twenty-one of these twenty-six remained carious at the second examination Three had been filled and two had been extracted. Of the teeth which had become carious, therefore, 19 2 per cent had received dental care Among the boys in the control group, twenty-five teeth had become carious, and of these, four had received dental care, or 16 per cent. In the case of the girls of the experimental group, 37.5 per cent of the teeth which became carious received dental care and of the girls in the control group 22.7 per cent of the teeth which had become carious received dental care. The per cent for boys and girls together in the case of the experimental group is 28 and the control group 19 2. While the numbers are small, the evidence points in the direction of the conclusion that the children of the experimental group received better dental care than did those of the control group. In neither group, however, can the instruction be regarded as having been very effective, since fewer than 30 per cent of the teeth which required dental care received it.

SUMMARY

The experiment was carried on in two fifth and two sixth grades of two public schools. Group 1 which saw the films was in one school and Group 2 which did not see the films was in the second school. The purpose was to measure the effect of seeing the motion pictures on the children's behavior.

The procedure consisted in giving the same instruction for thirteen school days to the parallel groups The teaching was done by a trained assistant. The subject-matter included instruction on the structure and development of the teeth, on diet, and the care of the teeth. The same detailed outline was followed with both groups With Group 1 oral instruction was supplemented by pictures, diagrams, and models, and also by two motion picture films entitled, Tommy Tucker's Tooth and Clara Cleans Her Teeth With Group 2 the motion picture film was not used. The number of children in Group 1 was 208 and in Group 2, 182.

The results of the instruction were measured by

1 An information test
2 A questionnaire in which the children reported on the care of their teeth at home and the food which they ate
3 Material brought in by the children
4. Condition of the teeth as determined by dental examination

1. *Information Test.*—The groups were first equalized on the basis of intelligence test scores Group 1 advanced from a score of 17.03 points to 29 47 points while Group 2 advanced from 14 85 points to 29 69 points. Group 1 thus advanced 12.44 points and Group 2, 14.84 points or 2.4 points more The film group thus gained less on the information test than the control group.

2 *Questionnaire.*—The children of both groups reported almost universal use of the toothbrush and eating of vegetables, fruit, and milk. Fewer children of Group 2 reported using the toothbrush daily (eighty-one per cent as compared with ninety-eight per cent). The difference seems hardly significant

3. *Material Brought In and Read*—Group 2 read about the same number of booklets and took home more literature, but Group 1 brought in many more clippings and contributed more original articles, cartoons, and pictures. This comparison is somewhat in favor of the film group

4 *Results of Dental Examination.*—Since age is a factor in the development and decay of the teeth, the groups were equalized by ages After such equalization the comparison was made of the average number of carious six-year-old molars of the two groups on examinations made before and after instruction In Group 1 the average number per child increased from 1.46 to 1.57 and in Group 2 the average number decreased from 1 90 to 1 80 The difference may have been due to a difference in the number of carious teeth on the first examination The groups were then equalized on the basis of the condition of their six-year-old molars on the first examination In Group 1 after equalization the average number per child of carious six-year-old molars decreased from 2.03 to 1 79, and in Group 2 the average number decreased from 2 02 to 1 87 This shows a slight advantage for the film group

A comparison was then made of the molars which were in perfect condition on the first examination A smaller number of perfect molars became carious in the case of Group 1 than in the case of Group 2, (20 6 per cent against 27.9 per cent) A larger number of these teeth received dental care in Group 1 than in Group 2, (28.0 per cent against 19 2 per cent) This comparison is slightly in favor of Group 1

The children who did not see the film apparently gained more information from the instruction than those that did The children who saw the film, however, appear to have been stimulated to somewhat greater activity in bringing in supplementary material After careful analysis has been made of the data, the group which saw the film appears to have improved somewhat more in the condition of their teeth than the pupils who did not see the film The difference, however, is not conclusive A slightly larger percentage of the group which saw the film received dental care, but fewer than 70 per cent of the defects of either group were corrected. Further experimentation would be desirable with somewhat more carefully controlled conditions and with films which may be more effective than those which were used in the experiment

INTELLIGENCE, MOTIVATION, AND ACHIEVEMENT

AUSTIN H. TURNEY

University of Kansas

Leading authorities in the field of psychology and of educational psychology recognize the existence of fundamental motivating mechanisms. Thus Gates writes of certain "cravings" and "urges"[1] and Woodworth of "dependable motives."[2] These fundamental motives are factors in school achievement to a greater or lesser degree, depending in part upon the method of teaching being used Thus we have many writers accepting the project method as psychologically sound because it enlists certain of these dependable motives [3] The literature of method is full of discussions concerning "motivation" or "motivating the pupil."

The beginning student in educational psychology sometimes fails to appreciate the direct and practical connection between fundamental motives and actual schoolroom work. It has been our experience that the importance of motivating mechanisms is made more clear if the treatment of this topic is linked up with the discrepancy between intelligence and achievement. This approach not only emphasizes the importance of the rôle of motivation, but also recognizes the fact that one great educational problem is that of determining the extent to which motivating mechanisms affect achievement, and of determining the possibility of affecting these mechanisms through teaching technique and choice of content.

A number of investigators have submitted data indicating that the discrepancy between intelligence test results and achievement usually found in schoolroom situations is not necessarily the result of faulty measures either of intelligence or of achievement On the contrary, this discrepancy is the natural result of other factors, chief among which are certain traits or types of behavior which for want of better terms we may call "industry," "persistence," "ambition," "school attitude," and "dependability."[1]

---

[1] Gates, A I "Psychology for Students of Education." The Macmillan Company, 1930, Chap VI

[2] Woodworth, R S "Psychology." Henry Holt and Co , 1929, Chap VI

[3] Jordan, A M "Educational Psychology " Henry Holt and Co , 1929, p 113.
Kilpatrick, W. H. "Foundations of Method." The Macmillan Company, 1926.

Douglass, H R "Modern Methods of High School Teaching " Houghton Mifflin, 1926, Chap XI.

This paper will present data suggesting that the behavior of pupils characterized by such terms as "industry," "perseverance," "dependability," and "ambition," as judged by teachers is probably an index of conditions of motivation, and indicating also the importance of the factors of motivation in school achievement. The data were secured in a study carried out in the University of Minnesota High School Every student in each of the four classes in the High School was rated by each of his teachers twice during the school year on each of nine traits The reliability of the judgments, as found by correlating first and second judgments made about sixty days apart, was, on the average, .80

Intercorrelations were calculated for the following variables [2]

1. Numerical value of all marks earned by each student during his entire residence in the high school
2. Chronological age
3. IQ derived from five or more tests
4. Mental age
5. Industry
6. Perseverance
7. Dependability
8. Ambition

For the Junior and Senior classes partial correlations of the first order were calculated holding constant in turn CA, MA, IQ, and marks.

Table I presents zero order correlations concerning the Junior class. Attention is especially called to the following groups of coefficients showing the correlations between certain variables:

Marks and MA	57	MA and industry	33
Marks and IQ	61	MA and perseverance	35
Marks and industry	80	MA and dependability	37
Marks and perseverance	80	MA and ambition	40
Marks and dependability	82	Industry and perseverance	96
Marks and ambition	84	Industry and dependability	96
IQ and industry	37	Industry and ambition	93
IQ and perseverance	48	Perseverance and dependability	95
IQ and dependability	.40	Perseverance and ambition	91
IQ and ambition	44	Dependability and ambition	93

---

[1] See Turney, A. H. "Factors Other than Intelligence That Affect Success in the High School as Indicated by Teachers' Marks" Minneapolis, Minn., University of Minnesota Press, 1930, pp. 31ff for a review of these studies

[2] The original data involve six other variables omitted as not pertinent to this analysis.

TABLE I.—INTERCORRELATIONS BETWEEN VARIABLES INDICATED
(*N* = 65 Juniors, University of Minnesota High School, 1926–1927)

	1 Marks	2 CA	3 IQ	4 MA	5 Industry	6 Perseverance	7 Dependability	8 Ambition
1. Marks		– 314	608	569	798	798	822	842
2 CA.			– 493	– 419	– 201	– 307	– 118	– 358
3. IQ				.952	365	481	397	444
4 MA					326	352	368	398
5 Industry						964	956	928
6 Perseverance							950	913
7 Dependability								934
8. Ambition								

Table II presents similar data concerning the Senior class in which attention is invited to the following correlations:

Marks and MA.	67	MA and industry	35
Marks and IQ	.69	MA and perseverance	31
Marks and industry	66	MA and dependability	41
Marks and perseverance	63	MA and ambition	39
Marks and dependability	72	Industry and perseverance	.96
Marks and ambition	68	Industry and dependability	90
IQ and industry	33	Industry and ambition	78
IQ and perseverance	.32	Perseverance and dependability	91
IQ and dependability	41	Perseverance and ambition	83
IQ and ambition	.40	Dependability and ambition	82

TABLE II—INTERCORRELATIONS BETWEEN VARIABLES INDICATED
(*N* = 48 Seniors, University of Minnesota High School, 1926–1927)

	1 Marks	2 CA	3 IQ	4 MA	5 Industry	6 Perseverance	7 Dependability	8 Ambition
1 Marks.		– 315	691	671	661	630	717	675
2 CA			– 736	–.712	– 255	– 202	– 265	– 390
3 IQ.				.986	330	315	407	399
4 MA.					348	305	405	389
5 Industry						956	895	780
6 Perseverance							911	834
8. Dependability								815
7 Ambition								

For our purposes we may summarize the two tables as showing a correlation between marks and IQ slightly higher than the average coefficient of correlation between measures of intelligence and measures of achievement reported in the literature.[1] Between marks and the traits of "industry," "perseverance," "dependability" and "ambition," the correlations are as high and, more often than not, higher than the correlation between IQ and marks or that between MA and marks The correlation between IQ and each of these four traits and between MA and each of these four traits is relatively low, but the correlation between each pair of these four traits is high

TABLE III —INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH CA CONSTANT
(N = 65 Juniors, University of Minnesota High School, 1926–1927)

	1 Marks	2 IQ	3 MA	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1. Marks		546	506	789	776	831	823
2. IQ			.942	311	308	392	330
3. MA				272	258	354	293
4 Industry					969	958	936
5 Perseverance						968	905
6 Dependability							963
7 Ambition							

The relatively low correlation between each of the traits and either IQ or MA suggests that neither are these "intellectual" traits nor are the judgments the effect of halo Halo effect of marks upon judgments of the traits can scarcely be present since each of the judges furnished only one-twelfth of the Junior marks and only one-sixteenth of the Senior marks.

In Tables III and IV are the partial coefficients of correlation between the variables with CA constant. Of the correlations especially pointed out above the greatest change is in the case of that between IQ and marks which is somewhat reduced In the main, the correlations between IQ and marks, IQ and the four traits, marks and the four traits, and between the traits themselves are not greatly affected when CA is held constant.

---

[1] Turney, A. H *Op cit*, Table I

Tables V and VI present the partial coefficients of correlations between each of the variables with IQ constant. It is to be noted that the correlations between marks and each of the four traits, "industry," "perseverance," "dependability" and "ambition," and also those between the traits themselves are but slightly changed.

TABLE IV—INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH CA CONSTANT
(N = 48 Seniors, University of Minnesota High School, 1926–1027)

	1 Marks	2 IQ	3 MA	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1. Marks		.721	670	635	610	695	633
2. IQ			977	219	252	327	181
3 MA				635	.610	695	633
4 Industry					956	880	766
5 Perseverance						900	837
6. Dependability							803
7. Ambition							

TABLE V—INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH IQ CONSTANT
(N = 65 Juniors, University of Minnesota High School, 1926–1927)

	1 Marks	2 CA	3 MA	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1 Marks		− 020	−.283	783	728	800	803
2. CA			184	026	− 092	098	− 176
3 MA				− 072	− 387	− 035	− 089
4 Industry					964	950	937
5 Perseverance						944	888
6 Dependability							921
7 Ambition							

The results of holding constant MA are shown in Tables VII and VIII The correlations between marks and the traits, and between the four traits themselves are but slightly changed

When marks are held constant, as shown in Tables IX and X, the correlations between the four traits, "industry," "perseverance,"

TABLE VI.—INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH IQ CONSTANT
(N = 48 Seniors, University of Minnesota High School, 1926–1927)

	1 Marks	2 CA	3 MA	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1 Marks		398	– 098	634	.601	661	.602
2 CA			147	– 019	047	057	– 156
3 MA				173	– 045	031	– 031
4 Industry					.953	884	.749
5 Perseverance						906	816
6. Dependability							757
7 Ambition							

"dependability" and "ambition," and both MA and IQ become very low and usually negative. The correlations between the traits themselves are still high

TABLE VII.—INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH MA CONSTANT
(N = 65 Juniors, University of Minnesota High School, 1926–1927)

	1 Marks	2 CA	3 IQ	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1. Marks		– 097	257	790	777	803	818
2 CA			– 333	– 075	– 188	043	– 229
3 IQ.				187	500	203	227
4. Industry					956	953	922
5 Perseverance						943	901
6. Dependability							925
7 Ambition.							

The foregoing data indicate that the four traits, "industry," "perseverance," "dependability" and "ambition" are more closely related to achievement than is either IQ or MA, and that they are not closely related to either IQ or MA, but that they are so closely related to each other as to show marked overlapping[1] In view of the high reliability found for all of these data we can only conclude that the teachers were in fact actually judging accurately the behavior of their students in a specific schoolroom situation where these traits represent the extent to which the students were motivated to conform to classroom require-

TABLE VIII.—INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH MA CONSTANT
(*N* = 48 Seniors, University of Minnesota High School, 1926–1927)

	1 Marks	2 CA	3 IQ	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1 Marks		523	286	614	602	657	605
2. CA			–.343	– 011	022	036	– 174
3 IQ				098	104	062	115
4. Industry					964	883	751
5 Perseverance						908	821
6 Dependability							786
7. Ambition							

TABLE IX.—INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH MARKS CONSTANT
(*N* = 65 Juniors, University of Minnesota High School, 1926–1927)

	1 CA	2 IQ	3 MA	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1. CA		– 401	– 307	088	– 098	257	– 182
2. IQ			931	– 252	– 008	– 227	– 158
3 MA				.260	– 207	– 213	– 181
4 Industry					908	877	784
5. Perseverance						860	739
6 Dependability							778
7. Ambition							

---

[1] Garrett, H E "Statistics in Psychology and Education" Longmans, Green & Co 1926, p 291

ments and to achieve in classroom activities. Hence these judgments represent largely the effects of motivation and they necessarily overlap as already shown. This overlapping of the traits and the lack of correlation between these traits and either IQ, MA, or CA point to the conclusion that the motivation existing in this situation was in reality the effect of "drive" or "dependable motives" as affected by the situation.

The significance and implication of these data are, we believe, rather great. They should serve to focus attention upon the real nature of the discrepancy between "intelligence" and school achievement. It seems clear that the two major factors in school achievement are intelligence on the one hand and "motivation" on the other. But motivation is not to be regarded as the product entirely of the immediate situation. The presence of fundamental driving mechanisms, the play of interests developed long before, and the varying effect of the immediate situation in the schoolroom are probably all a part of what we call "motivation."

TABLE X.—INTERCORRELATIONS BETWEEN VARIABLES INDICATED WITH MARKS CONSTANT

(N = 48 Seniors, University of Minnesota High School, 1926–1927)

	1 CA	2 IQ	3 MA	4 Industry	5 Perseverance	6 Dependability	7 Ambition
1 CA		− 026	− 712	066	− 007	− 059	− 255
2 IQ			972	223	− 211	− 175	− 126
3 MA				− 172	− 204	− 148	− 118
4 Industry					943	808	606
5. Perseverance						851	719
6 Dependability							650
7 Ambition							

Our data serve only to emphasize the presence and importance of factors of motivation. It would be desirable to know to what extent these factors are the result of specific environmental situations such as the classroom. To what extent can the "dependable" motives be influenced by specific schoolroom situations such as the use of the project method or the socialized recitation? In other words to what extent can the industry and ambition of pupils be changed through educational influences? The variability in responsiveness to specific stimulating situations in the classroom may be as much a biologically fixed quality as "intelligence" is usually considered to be.[1] On the other hand, it may be vastly more subject to educational (*i e*, environmental) influences

If we accept for purposes of discussion the concept of *g* as set forth by Spearman[2] the situation can be presented in the following way Two students possessing equal amounts of *g* would not necessarily utilize it equally in the same classroom activity Moreover, even objectively measured, achievement must involve a great deal more than *g*. Not all classroom activities would correlate highly with measures of *g*.[3] Many of them would depend as much or more upon just these characteristics of the student which we label "industry," or "perseverance" for success. Hence the correlations shown between "industry," "perseverance," "dependability" and "ambition," and marks, are probably measures of a true relationship To the extent that the classroom work involves activities not largely dependent upon a pupil's *g*, or to the extent that activities very largely involving *g* are not attended to in equal proportion by students approximately equally equipped, a teacher's marks could not correlate perfectly with measures of intelligence.

These facts might well be borne in mind in using mental tests. They do not minimize the utility of tests of "intelligence" but rather enhance their value. In ability grouping, for example, until the possibilities of motivation are known, we must agree with Billett,[4] who is one of the few to point out the advisability of grouping on measures of mental ability alone. To include in the criteria for grouping measures of past achievement covers up the very important effects of motivation The possibilities of educational guidance, and the understanding of difficulties are much greater with the use of the single criterion If criteria such as achievement tests which measure, in part, the effects of motivation are used, they should be utilized in a manner permitting the teacher's knowledge of their presence and weight.

---

[1] I am not unaware of the arguments against this concept, as in the Twenty-seventh Yearbook of the National Society for the Study of Education, Part I, but see Jennings, H S "The Biological Basis of Human Nature" New York W. W Norton, 1930, p. 24

[2] Spearman, C. "The Abilities of Man" The Macmillan Co, 1927.

[3] Ibid., Chap. XII.

[4] Billett, R O A Controlled Experiment to Determine the Advantages of Homogeneous Grouping Ohio State University, *Educational Research Bulletin*, Vol VII; Nos. 7, 8 and 9, April 4, 18, May 2, 1928

THE "OMISSION" AS A SPECIFIC DETERMINER IN THE TRUE-FALSE EXAMINATION

C. C WEIDEMANN

Teachers College, University of Nebraska

Do students tend to omit items of the true-false examination that are *true* more often than they tend to omit items that are *false?* In this study each item is either true or false by published authority as the criterion

The *specific determiner*[1] is a word or phase which occurs in the *external form* of a test item and alters the *internal constitutiom* of that test item so that the correct response is either more often *true* than *false* or more often *false* than *true.* The *specific determiner* in any given test item may be so controlled by the individual who constructs the true-false examination that the correct response is approximately either as often *true* as *false* or as often *false* as *true*

Wood[2] differentiates between *internal constitution* and the *external form* of examination items He says.

> The essential character of a question, and consequently its "statistical behavior" depends upon its content as well as its external form The content of a question is often very subtle; many of the "simplest" questions are found on inspection to be compounds of many subtleties. It seems quite reasonable to suppose that within the limits of accepted forms of questions, the "inner" nature of a question is more important than its "outer" form; although as a matter of common sense, one may say that the "outer" aspect exerts more influence as we go down the intellectual or educational ladder and less as we ascend it

Assume that the *specific determiners* of a true-false examination have been controlled, yet the variable of the temperament of the examinees is still present. The temperament of examinees may cause some to respond more often *true* than *false* and other to respond more often *false* than *true* to the same test item in a given true-false examination This variable of temperament of the examinee seems to be somewhat a specific determiner in relation to any given test item A study of this point needs to be made

---

[1] Weidemann, C C. "How to Construct the True-false Examination" Bureau of Publications, Teachers College, Columbia University, 1926, p 68

[2] Wood, Ben D Studies in Achievement Tests *Journal of Educational Psychology,* Vol XVII, January, February, 1926, pp 1–22, 125–139

So far as an individual examinee is concerned, no direct technique seems to be available to control the temperament of that examinee while responding to the items in true-false examinations whose *specific determiners so far as external form are concerned*, have been controlled on the assumption that *score equals right minus wrong*. In the case of large groups of examinees temperament might be considered to be a chance factor, compensating in such a way that the tendency for some examinees to respond more often *true* than *false* would be offset by the tendency of others to respond more often *false* than *true*. This study assumes the factor of temperament of the examinees to be a chance factor.

Fritz[1] and a few others have claimed that students tend to respond to true-false items more times as *true* than as *false* The figures are sixty-two per cent *true* and thirty-eight per cent *false*, with technical information quite unknown to the students He reports practically the same result with information which was studied by the students as a regular part of the class work.

Ruch[2] reports "a number of studies similar to that of Fritz" and gives fifty-two per cent marked as *true* and forty-eight per cent marked as *false*. In an unpublished study by D. D. Durrell, and C. L Cushman, (data available through G. M. Ruch, University of California) "the results indicate that pure guessing, when it occurs, is roughly a 50:50 situation."

Rutledge[3] found "that when the 'same' items were stated in both true and false forms, they were equally difficult "

These three studies make no mention of the "omission" as a specific determiner

In the fall of 1926 in a course entitled "Foundations of Modern Education," the writer used seven true-false examination tests based upon F. P Graves's, "A Student's History of Education " At the beginning of the semester, the class totalled three hundred thirty-eight divided into three sections of ninty-nine, one hundred three, and one hundred thirty-six students respectively. A total of three hundred thirty-one students included in this study remained throughout the semester. The members in each section were given a reading assignment from Graves and within two weeks were given a true-false test based upon the reading assignment The object was primarily to measure immediate retention rather than delayed retention. No test contained any items which aimed to review material covered in any previously given test

---

[1] Fritz, M F Guessing in a True-false Test. *Journal of Educational Psychology*, Vol XVIII, 1927, pp. 558–561

[2] Ruch, G. M "The Objective or New-type Examination " Scott, Foresman and Company, New York, 1929, pp. 365

[3] Rutledge, R E "The True-false Examination in Elementary Psychology, with Suggestions for Its Improvement " Unpublished Ph.D Thesis, University of California, 1927

In order to reduce cheating to a minimum each of the seven bi-weekly tests was administered in either Forms A, B and C or Forms A and B. The students seated in Columns 1, 2, 3 had Forms A, B and C respectively; Columns 4, 5 and 6 had Forms A, B and C respectively, and so on After the forms were answered they were collected by the instructor Then Columns 1, 2 and 3 had Forms B, C, and A respectively, Columns 4, 5 and 6 had Forms B, C and A respectively, and so on. After the second order of Forms were answered they were collected by the instructor. Then Columns 1, 2, and 3 had Forms C, A and B respectively, and Columns 4, 5, 6 had Forms C, A, and B respectively, and so on A similar procedure was followed when the bi-weekly test consisted of Forms A and B. The total test time of each bi-weekly test was from twenty-five to forty minutes

The frequency with which each item of each Form of a given bi-weekly test was omitted was tabulated for each student taking each test All results are reported upon the basis that the number of *true* items equals the number of *false* items in each of the seven bi-weekly tests The *specific determiners* of "all and never," "more than, less than," etc, and "If . then" were controlled on the assumption that the *score equals rights minus wrongs*

The directions commanded the student to mark with a plus (+) only those statements which were entirely *true,* if the statement were partly or entirely *false* it was marked with a zero (0). A response space was provided to the left of each statement The students were told *not* to guess

Table I displays the number of times each omitted item was either *true* or *false* A grand total of 3064 students responded to a grand total of 495 true-false statements, accumulating a grand total of 3954 omissions, in seven bi-weekly tests administered during one semester Table I shows that forty-four per cent of the omitted statements were *true* and fifty-six per cent of the omitted statements were *false* Is the twelve per cent difference significant? The critical ratio value is sixteen, so that the difference is significant

TABLE I.—NUMBER OF TIMES *Omitted* ITEMS ARE EITHER *True* OR *False* FOR EACH OF THE SEVEN TRUE-FALSE TESTS

Bi-weekly true-false test number	The number of times test items were omitted by the student, which items were		Total omissions
	True	False	
1	2	3	4
1	737	655	1392
2	270	283	553
3	159	400	559
4	219	203	422
5	57	46	103
6	112	229	341
7	172	412	584
Totals	1726	2228	3954
Per cent	44	56	100

Even though the statistical per cent difference is significant, there is relatively little value in the per cent difference so far as affecting the *score and placement of an individual student is concerned*

Assume:

1 The score equal to the total number of items minus the *omissions* minus two times the wrongs, *i.e.*, $S = T - O - 2W$.

2 That on the average not more than twenty items out of every one-hundred items of a true-false examination would be omitted by a given student Some may omit more and others less items.

3 *That every item so omitted would then be marked false*

Then upon the basis of this study, fifty-six out of every one hundred omitted items arbitrarily marked false would be *right* and forty-four would be *wrong*. $S = 56 - 44 = 12$ But, it is assumed that on the average, not more than twenty items out of every one hundred items of a true-false examination would be omitted and arbitrarily marked *false.* Thus the contribution of the *omission* 12 × 20 per cent or 2.4 points of score. This change of score seems administratively to be relatively insignificant in the degree to which it may affect the placement of an individual in a group

1. Three hundred thirty-one students who took the seven bi-weekly true-false tests based upon Graves' "A Student's History of Educa-

tion," in crude per cents omitted items that were *false* more often that they omitted items that were *true.*

2. The *omission* in the seven bi-weekly true-false examinations of this study is a statistically significant *specific determiner.* The critical ratio value is 16

3. It seems to the writer that in a very carefully constructed true-false examination "omissions" are probably as often *true* as *false,* and as often *false* as *true*

4. From the standpoint of affecting the grade placement of an individual student in a group, this study would tend to indicate that the *omission* as a specific determiner is not significant

5 Further study will probably show the *omission* of the true-false examination to be statistically insignificant.

FACTORS INVOLVED IN CHILDREN'S FRIENDSHIPS[1]

GLADYS GARDNER JENKINS

University of Chicago

The following study was made in the effort to discover definite factors which influence the forming of friendships among children. There are, of course, many factors of a personal nature not yet possible of analysis, but there are others such as the relative intelligence of child and friend, age differences, play interests, and like factors which are measurable in some degree and may bear upon the question The study included two hundred eighty boys and girls representing a cross section of the junior high schools of Riverside, California, a city of approximately 35,000 inhabitants.

CHRONOLOGICAL AGE

The chronological age of one hundred eighty children was compared with that of their friends. The range for one hundred twenty-five girls was twelve years and five months to sixteen years and eleven months; for their friends twelve years to nineteen years and eight months. For fifty-five boys the age range was twelve years and three months to sixteen years and four months; for their friends twelve years and five months to seventeen years The mean chronological age for the children was fourteen years and three months and for their friends fourteen years and five months, with a standard deviation of 10.6 months for the children, and 13 0 months for their friends.

Of the one hundred eighty boys and girls studied ninety-seven made their friendships in school, eighty-three outside of school. The mean chronological age for those friendships made in school was fourteen years and four months for both the children and their friends, with a standard deviation of 10.3 months for the children and eleven months for their friends. The mean chronological age for the eighty-three friendships made outside of school was fourteen years and two months for the children and fourteen years and six months for their friends, with a standard deviation of 10.9 months for the children and 16 1 months for their friends. As there is a larger spread in the chronological age range for those friends which are made outside of school than for those made within the school, it would seem as if the school has a tendency to diminish the age difference in children's friendships. In spite of the larger standard deviation for the friendships made outside of school there is no uniform or perceptible tendency for children to choose friends either older or younger than themselves The data are presented in Table I.

---

[1] The writer wishes to express her indebtedness to the guidance of Dr Frank N. Freeman in this study, and to former Superintendent A N Wheelock and Principal F P Taylor of the Riverside School System for their cooperation in making available their records

TABLE I—COMPARISON OF CHRONOLOGICAL AGE DIFFERENCE BETWEEN CHILD AND FRIEND WITH THE PLACE WHERE THEY MET

Age difference, months	Neighborhood and places other than school		School	
		Per cent		Per cent
Within one month	6	9.6	5	4.8
1 to 6	22	31.0	40	38.8
7 to 12	17	24.7	46	44.7
13 to 18	12	17.5	6	5.9
19 to 24	4	5.8	3	2.9
25 to 30	4	5.8	3	2.9
31 to 36	1	1.4		
37 to 42	2	2.9		
43 to 48				
49 to 54				
55 to 60	1	1.4		

The Pearson product-moment correlation coefficient was +.470 ± .039 The correlation coefficient for the ninety-seven friendships made within the school was +.529 ± .049, for the eighty-three friendships made outside of school, +.462 ± .073 It would therefore seem that the tendency for children to choose friends of their own age exists independently of school grouping although it may be increased by it

INTELLIGENCE QUOTIENTS

Intelligence quotients obtained by the Terman Group tests within the two previous years were available from the school records for one hundred ninety-seven of the children, one hundred thirty-four girls and sixty-three boys The correlation coefficient was +.332 ± .043 as against a correlation coefficient of + 470 ± .039 for the chronological age. The correlation coefficient for the intelligence quotients of one hundred eight friendships made within the school was + 299 ± 059, for eighty-nine friendships made outside of the school the correlation coefficient was +.374 ± .062. Inasmuch as the correlation coefficient for the friendships made out of school is slightly higher than that for those made in school, even allowing for the equalization which might occur because of the probable error, the indications would seem to be that intelligence quotient is an independent factor in children's friendships, not simply the outcome of school sectioning.

The mean intelligence quotient for both the children and their friends was 109 8 with a standard deviation for both groups of 15.3 For those friendships made in school the mean intelligence quotient is 109 5 for the children, 111 1 for their friends, with a standard deviation of 14.1 for the children and 14 8 for their friends. For those friendships made outside of school the mean intelligence quotient is 110 1 for the children, 108 2 for their friends, with a standard deviation of 16 7 for the children and 15.7 for their friends.

MENTAL AGE

The mental age of one hundred seventy-two children and their friends, of which one hundred twenty were girls, fifty-two boys, was computed for the time of the study rather than for the time at which the friendship was made, as it did not seem possible to secure an accurate record of the exact time when each friendship was formed. Inasmuch as all the children studied had been tested by the Terman Group test within the two previous years it was felt for the purpose of this study it would be satisfactory to assume that the intelligence quotients had remained constant and to compute the mental age from the intelligence quotient and the chronological age at the time of the study.

The mental age range for both the children and their friends was eleven years and one month to twenty-one years eleven months. Seven children had the same mental age as their friends, thirty-seven children chose friends with a mental age within six months of their own, sixty-two had a mental age more than six months higher than their friends, sixty-six more than six months lower There seems to be an equal tendency for a child to choose a friend with a higher mental age as there is for him to choose a friend with a lower one. The mean mental age for the children was fifteen years and seven months with a standard deviation of 22 7 months, the mean mental age for their friends was approximately the same, fifteen years and ten months with a standard deviation of 21 9 months. The correlation coefficient was + 423 ± .042.

A comparison of the correlation coefficients for chronological age, mental age, and intelligence quotient reveals the following data in order of the probable importance.

TABLE II —A COMPARISON OF THE CORRELATION COEFFICIENTS OF CA, MA AND IQ

Subject	Correlation coefficient	PE
Chronological age	+ 470	± 039
Mental age	+ 423	± 042
Intelligence quotient	+ 332	± 043

SCHOOL DIVISIONS

The fact that children seem to tend toward a choice of friends of approximately the same chronological age and mental age would suggest that these friends might also tend to be in the same grade or section in school. The following facts bear out this supposition. Of two hundred fifty-four children studied (one hundred twelve boys, one hundred fifty-one girls) one hundred thirteen of their friends were in the same section, sixty-two in the same grade but not the same section, seventy-nine in other grades. Of the one hundred twelve boys twenty-three or twenty per cent chose friends in a higher grade, twenty-one or eighteen per cent in a lower grade, and seventy or sixty-one per cent in the same grade. Of the one hundred fifty-one girls eight or five per cent chose friends in a lower grade, thirty-two or twenty-one per cent in a higher grade, one hundred eleven or seventy-four per cent in the same grade The boys seem to have a higher range of friendships than the girls, sixty-one per cent of the friends of the boys were in the same grade as against seventy-four per cent of the girls This tendency, if it holds true with a larger group, may be due to the fact that there is a slight tendency for boys to make more friendships in the neighborhood and through clubs than girls and, as has been shown, the age range for friendships made outside of school tends to be slightly higher than for school-friendships.

PLACES IN WHICH THE FRIENDSHIPS DEVELOPED

In order to discover whether the school, the neighborhood, the club, or other environmental factors have the most effect upon the making of friendships the following factors concerning the places where the friendships were made were investigated.

1. School
2. Neighborhood
3. Club—Girl Scouts, Boy Scouts, Y M. C A , Girl Reserves
4. Church.
5. Miscellaneous—family contacts, special interests—music, etc

Information was obtained from two hundred fifty-five children (one hundred sixty girls, ninety-five boys). Table III indicates the number of friendships made in each of these places

TABLE III.—NUMBER OF FRIENDSHIPS IN RELATION TO THE PLACE WHERE THE CHILDREN MET

Place	Number of friends	Per cent
School	138	54
Neighborhood	63	25
Church	20	8
Club	8	3
Miscellaneous	26	10

The table shows that the school was the place in which the largest number of friendships was made, with the neighborhood ranking second. A slight tendency was found for the boys to make more of their friendships in the neighborhood and places other than school than for the girls to do so. Forty-nine per cent of the boys and forty-four per cent of the girls made their friendships in places other than school

Table IV indicates that the section grouping has a distinct influence upon the friendships which children make within the school, whereas for those friendships made out of the school the section has very little significance

TABLE IV.—RELATION OF SECTION GROUPING TO FRIENDSHIPS MADE IN THE SCHOOL AND OUT OF THE SCHOOL

School division of friend	Friendship made in school, per cent	Friendship made out of school, per cent
Same section	64	19
Same grade (not same section)	25	21
Other grade	11	57

OCCUPATION OF PARENT

In order to attempt to determine the influence of the social-economic status of families upon the friendships of the children a comparison was made between the occupation of the father of each child with that of the father of his friend For the purpose of the study the occupations were grouped into the following four groups, coinciding approximately with the social groupings of the community:

1 Unskilled and semi-skilled labor

2. Skilled labor (carpenter, electrician, etc).

3 Semi-professional (nurse, school teacher, etc.), commercial, owner of small ranch or citrus grove, clerical.

4. Professional (engineer, physician, etc), business executive, owner of large ranch or citrus grove.

Two hundred thirty-seven cases were studied (ninety-five boys, one hundred forty-two girls). It must be remembered that California is a large middle class state, and therefore the total distribution of children is largely weighted toward the skilled labor and commercial groups. Table V indicates the results secured

TABLE V.—COMPARISON OF OCCUPATION OF PARENT OF CHILD WITH PARENT OF FRIEND

Parent's occupation		Unskilled		Skilled, etc		Commercial, etc		Professional, etc	
Total number of cases	Per cent of total number	Number	Per cent	Number	Per cent	Number	Per cent	Number	Per cent
Unskilled labor									
16	7	11	69	4	25	1	6		
Skilled labor, etc									
121	51	4	3	83	69	29	24	5	4
Commercial, etc									
71	29			11	16	50	71	9	13
Professional, etc									
30	13			1	3	7	23	22	74

These figures would seem to indicate that the social-economic status of the parents in the community is an important criterion in the choice of friends which the child makes. In each case the number of friends chosen from the same social-economic group equals more than sixty per cent of the total distribution.

	Per Cent
Unskilled labor	69
Skilled labor, etc	69
Commercial, etc	71
Professional, etc	74

The correlation coefficient +.716 ± .032 strongly expresses this tendency for children to choose friends of the same social-economic group As there were only four groups the correlation coefficient was corrected for coarse grouping by Shepherd's formula with a resulting correlation coefficient of +.817 The correlation coefficients for the boys and their friends and for the girls and their friends were practically the same, +.695 ± 036 for the boy group, and +.724 ± .027 for the girl group.

It seems probable that the significance of the high correlation coefficient is not greatly influenced by the proximity of the homes of the children and their friends, as only twenty-five per cent of the children stated that their friendships had been formed in the neighborhood. It might also be mentioned that the clubs and churches in Riverside are centralized rather than located in special residential districts, so that it is probable that those friendships made through these agencies were not unduly influenced by neighborhood divisions. There is also a suggestion of a tendency for children to reach into a higher social-economic group in the choice of a friend, forty-eight children chose friends in a higher group against twenty-three who chose them in a lower one. The facts given show, however, that the stronger tendency in a child's choice of a friend is toward those of the same social-economic level.

Play Interests

Lehman's Play Quiz was given to one hundred twenty-six children (twenty-eight boys, ninety-eight girls) and their friends. Unfortunately the test was given one week after the other data had been collected, and in the meantime a rearrangement of school groups had made it impossible to reach all the boys who had been studied at the original time. The number of boys, therefore, is too small to permit any definite statement concerning the findings but the trends will be of interest In scoring the tests the number of activities carried on in the past week because the child liked to do them were checked against those of his friend, and the number of identical interests recorded. The child's interests were then checked against those of cases picked at random from the children studied, three cases for each of the girls, six for each of the boys. The mean number of like interests for child and friend, and child and other than friend is shown in Table VI. It would seem as if there were a slight tendency for the children and their friends to have the larger number of common interests.

TABLE VI.—MEAN NUMBER OF LIKE INTERESTS FOR CHILD AND FRIEND, AND CHILD AND OTHER THAN FRIEND

CHILDREN	MEAN NUMBER OF LIKE INTERESTS
Boys	
Boy and Friend	18.5
Boy and other than friend	16
Girls	
Girl and Friend	16.5
Girl and other than friend	14.5
Boys and Girls.	
Child and friend	17
Child and other than friend	15

A study was next made of Section D of the test in order to discover whether the child and best friend tend to have primary interests in common. Table VII shows that more children had a common interest in one of the three things they like to do best with their friend than with the children picked at random from the group. Twenty-seven per cent of the children studied had one of their three greatest interests in common with their best friend as compared with nineteen per cent who had one of their greatest interests in common with children selected at random.

TABLE VII.—A COMPARISON OF COMMON PRIMARY INTERESTS BETWEEN CHILD AND FRIEND, AND CHILD AND OTHER THAN FRIEND

Number of primary interests	Per cent of common primary interests between child and friend	Per cent of common interests between child and other than friend
1	27	19
2	5	1
3	1	

SUMMARY OF CONCLUSIONS

Social-economic Status of Parents—Of the factors considered in this study primary importance is assigned to the social-economic status of the parents. The correlation coefficient for the social-economic position of parent of child and parent of friend was +.716 ± 032 Corrected for coarse grouping by Shepherd's formula the coefficient was +.817. This high correlation does not seem to have been significantly influenced by the proximity of homes as only twenty-five per cent of the total number of children stated that they made their friends in the neighborhood There was also a suggestion of a tendency for the children to choose friends of a higher social-economic group; forty-eight had friends in a higher group, twenty-three in a lower.

Chronological Age.—The correlation coefficient for chronological age was +.470 ± .030. This is probably somewhat greater in those friendships made in school Children tend to choose friends within one year of their age. There is no tendency to choose friends either older or younger. Friendships made in the neighborhood have a larger age range than friendships made in school.

Mental Age and Intelligence Quotient—The correlation coefficient for mental age was +.423 ± 042. The correlation coefficient for intelligence quotient was +.332 ± 043 The fact that the intelligence quotient correlation coefficient does not seem to be influenced by school groupings would suggest a spontaneous tendency to choose friends of the same approximate intelligence

School Divisions and Place of Meeting—There is an almost equal division in the number of friendships made in the school and those made in the neighborhood or through home contacts, churches, and clubs It must be recognized, however, that the school is the greatest single source with the neighborhood ranking second with twenty-five per cent of the total number of friendships. Of those friendships made in school the grade-section divisions seem to be of importance.

Play Interests—According to the data secured from Lehman's Play Quiz there seems to be a slight tendency for children to have a greater number of like interests with their best friends than with other children. It can not be stated as to whether this common interest was the cause or outcome of the friendship.

A STUDY OF CLASSROOM BEHAVIOR

WILLARD C. OLSON

University of Michigan

Considerable interest has been shown in recent years in the development of techniques for the quantitative expression of personality traits not easily amenable to measurement by test methods. By the use of carefully defined categories of overt behavior, relatively short time samples, repeated observations, and systematic recording, the writer has attempted to accomplish the task of measurement through direct observation.[1] The adaptability of the general method to a variety of behavior traits has been illustrated by Goodenough.[2] The present investigation was initiated to test further the applicability of the method and to acquaint a class in the psychology of personality with its use.

Whispering was chosen as the behavior to be observed. Wickman[3] has shown that whispering ranks first in frequency of occurrence among behavior problems in children, although teachers do not regard its occurrence in a child as a serious behavior disorder. A behavior problem is defined frequently as a discrepancy between the environmental demands made on an individual and his adjustment to them. Whispering, according to this definition, may or may not be a behavior problem in a particular classroom depending upon the requirements made with respect to it. The present report is concerned with an attempt to measure and describe whispering behavior in schoolroom situations quite apart from any judgment concerning its desirability or undesirability.

COLLECTION OF DATA

Arrangements were made to permit the class to spend one hour, from 9:00 to 10.00 A. M., in observation in an elementary school.[4]

---

[1] Olson, Willard C. "The Measurement of Nervous Habits in Normal Children." Minneapolis: The University of Minnesota Press, 1929.

[2] Goodenough, Florence L. Measuring Behavior Traits by Means of Repeated Short Samples. *Journal of Juvenile Research*, Vol. XII, 1928, pp. 230-235.

[3] Wickman, E. K. "Children's Behavior and Teachers' Attitudes." New York: The Commonwealth Fund Division of Publications, 1928.

[4] The writer is indebted to Miss Katherine G. Young, Principal of the Marcy School, Minneapolis, for cooperation in this project.

Each observer was supplied with a mimeographed sheet of instructions giving explicit directions for making the observations[1] A whisper for the purpose of the study was defined as an unauthorized oral communication—whispered or aloud—to a person or persons other than the teacher The unit of measurement was defined as one or more acts of whispering per five-minute period. Two observers were stationed in each room at opposite sides and all rooms in the building studied at the same hour. One student substituted a series of observations made on a twelfth grade class in economics. Ten successive observations were made and recorded in a prescribed manner on seating charts of the rooms The total number of time samples in which a child was recorded as whispering was called his *whisper score.*

RELIABILITY OF WHISPER SCORES

The reliability of the total score for ten five-minute records was obtained by correlating the values obtained on each child by the two observers in each room The coefficient of correlation between the two sets of scores varied between 20 and .78. The predicted reliabilities (Spearman-Brown formula) for twenty five-minute periods varied between 33 and 88 for the grade groups studied (Table I) The reliability of the single observer of a high school class was studied by correlating the sum of the odd numbered observations with the sum of the even numbered observations. The reliability coefficient for five observations was found to be .67 with .80 the predicted value for the ten observations.

In the room where the lowest reliability coefficient was found, the method of teaching called for frequent shifts in the seating arrangement Since the observers did not know the children except as a seat location, such changes introduced difficulties. A greater amount of time would have to be spent in such a situation, the room program slightly modified, or the children identified as individuals, to secure the requisite reliability It is probable that some initial training in the use of the method would improve the accuracy of the observers Increased reliability would be obtained by making a larger number of observations

It was felt that the records obtained showed sufficient consistency to warrant further inquiry into their distribution and significance In the subsequent analysis, the average of the records of two observers was taken as the score for each child in the elementary school. The high school records are the results for a single observer.

---

[1] Copies of the instruction sheet may be obtained from the writer upon request.

Table I.—Reliability of Whisper Scores

Group	*N*	r, 10 obs.	r (Sp. Br.), 20 obs.
1*B*	32	20	33
3*B*–2*A*	35	78	88
3*A*–3*B*	34	76	86
4*A*–4*B*	40	60	75
5*B*–4*A*	35	43	60
6*B*	31	29	45
6*A*	30	35	52
12	27	80	89

Distribution of Whisper Scores

Under the conditions of the method, the child who whispered in each of ten five-minute intervals would have a score of ten, while the child who did not whisper at all would have a score of zero. Scores for the elementary grades (Table II) varied between zero and eight, with a mean manifestation of 2.2. Since the score for each child in the elementary school is the average of records from two observers, the variabilities of the scores are somewhat curtailed. An examination of the table reveals the presence of a large number of zero values. In a sense, the method fails to differentiate about 33 per cent of the elementary grade children. Previous experience with the method in the measurement of other types of behavior suggests that added observations are needed in order to further differentiate the zero scores. The scores for the single observer of the twelfth grade vary between zero and ten with a mean of 4.5 and relatively few zero scores.

No reliable sex differences appear in the table. Although the data were sexed for purposes of analysis, the lack of any constant trend has led the writer to deal with only the totals in the remainder of this report.

Relation of Whisper Scores to Grade in School

No clear-cut relationship between the mean amount of whispering and the grade in school can be demonstrated with the number of cases available (Table III). The data suggest a larger amount in the first grade than in subsequent grades in the elementary school. On the other hand, a previous table (II) shows an almost equal amount in a group of high school seniors. It would appear that the gross amount of whispering in a particular grade is a function of the situation rather than the reflection of any fundamental developmental continuum. Deviation from the mean of a given situation is probably of more significance than the raw score. Such fluctuations as are found from room to room cannot be attributed solely to errors of measurement for two observers tend to give similar descriptions (Table IV).

TABLE II.—DISTRIBUTION OF WHISPER SCORES BY SEX

Scores	Elementary grades[1]			12th grade		
	Boys	*Girls*	*Total*	*Boys*	*Girls*	*Total*
10					2	2
9					2	2
8	1		1			
7	1	4	5	1		1
6	2	3	5	1		1
5	6	4	10	3	1	4
4	6	5	11	3	1	4
3	19	13	32	5		5
2	15	13	28	2		2
1	27	30	57	1	1	2
0	32	43	75	1	3	4
N	100	115	224	17	10	27
Md	1.8	1.5	1.6	3.9	5.0	4.1
M	2.3	2.1	2.2	4.0	5.3	4.5
SD_{dis}	1.8	1.9	1.8	1.7	4.2	3.0

[1] Cases without age records omitted from this tabulation.

WHISPERING, INTELLIGENCE, AND SCHOOL ACHIEVEMENT

Intelligence quotients, based on a single group test, were available for children in the 6A grade. The coefficient of correlation between whisper scores and intelligence quotients was –.34 ± .11. Intelligence quotients, based on the mean of five group tests, and honor-point ratios, based on all marks for three years of high school work, were available for the small group of high school seniors. The intercorrelations are presented in Table V. The relationship of whispering to honor-point ratios (–.30 ± .12) is higher than that with intelligence quotients (–.17 ± .13). Data on larger groups are necessary to establish these trends with certainty. The evidence available suggests that whispering is not simply an adventitious circumstance, but is a reflection in part of adjustment to the classroom situation. The more intelligent and scholarly, interested in the work of the class, adjust to it and whisper less. It is conceivable that the relationship would change under varying conditions of work and interest.

WHISPERING AND MARKS IN CONDUCT

In a sense, the validity of the whisper scores is intrinsic and reliability and validity coefficients become synonymous for the narrowly defined behavior which is the subject of study. If now a claim is made that the whisper scores are symptomatic of broader categories of behavior, external criteria must be used to evaluate the method.

TABLE III.—WHISPER SCORES BY GRADES IN AN ELEMENTARY SCHOOL[1]

	Grades						Total
	1B	*2A*	*3B–3A*	*4B–4A*	*5B*	*6B–6A*	
N	30	19	45	40	22	59	224
Md	4.6	1.6	0.8	1.8	0.8	1.3	1.6
M	4.8	2.5	1.4	2.0	1.2	1.8	2.2
SD_{dis}	1.5	2.0	1.6	1.4	1.0	1.3	1.8

[1] Cases without age records omitted from this tabulation.

TABLE IV.—COMPARISON OF MEAN SCORES AND VARIABILITY OF TWO OBSERVERS IN EACH CLASSROOM

Room	*N*	Mean		SD_{dis}	
		Observer 1	Observer 2	Observer 1	Observer 2
1B	32	5.6	1.4	2.0	2.0
2A–3B	35	2.9	2.1	2.7	1.8
4B–4A	40	2.6	2.5	1.5	1.4
6A	30	2.6	2.5	1.0	1.9
6B	31	1.9	1.2	1.2	9
3B–3A	34	1.4	1.2	1.7	9
4A–5B	35	1.4	1.0	1.2	7

Children in the elementary school studied were given marks in conduct by teachers in grades three and above. The marks were in terms of letter ratings, *A*, *B*, *C*, *D*, and *E*. The summary for a semester was taken as the mark for each child. There is a suggestion that the significance of whispering as one of the criteria of conduct increases as one goes from the freedom of the early school years to the more formal procedures of the upper grades (Table VI)

TABLE V —RELATIONSHIP BETWEEN WHISPER SCORES AND OTHER VARIABLES IN A HIGH SCHOOL CLASS ($N = 24$), PER CENT

	IQ	HP ave.	CA
Whisper	− 17	− 30	18
IQ		68	− 55
HP Ave			− 31

TABLE VI —THE RELATION OF WHISPER SCORES TO MARKS IN CONDUCT IN SUCCESSIVE GRADES

Grade	*N*	*r*
6*B* .	20	− 60
5*A*–5*B*	30	− 30
4*B*–4*A*	36	− 08
3*A*–4*B*	30	10

SUMMARY

A group of students in a course in the psychology of personality collected data on the occurrence of whispering in elementary and high school children by the use of a time sampling technique. Each student made ten consecutive five-minute observations The reliability coefficients for the average of two unpracticed observers varied between .33 and 88 in the rooms studied. No significant sex differences appeared in the mean whisper scores. The amount of whispering present in the various grades appeared to be a function of the situation rather than the reflection of a developmental trend The evidence suggested that the more intelligent and scholarly whispered somewhat less than other children. The significance of whispering as one of the criteria entering into marks of conduct increased with the grades.

REPLY TO PROFESSOR KELLEY

KARL J. HOLZINGER

University of Chicago

In the May issue of this Journal Professor Kelley has taken objection to some comments I have made[1] regarding an example taken from his book, "Crossroads in the Mind of Man."

To save print I refer the reader to the data in the papers cited above. In my March paper I tried to do the following things:

1. Show that Thorndike's CAVD intelligence test has a common factor with three other tests and determine its correlation with this factor: ($r_{1g} = .96$).

2. As an example that a "parsimonious" explanation of correlations is desirable I took four out of nine tests from some of Professor Kelley's data and found one common factor an adequate explanation. I then gave Professor Kelleys' pattern based on these four tests and five others. This pattern was very elaborate, having one common factor, three substantial group factors and several specific factors. On p. 164 of my article I say,

> The above example is but a *fragment* of Professor Kelley's work on these data and is not included by way of criticism, but merely because the numeral work was at hand. As far as these *four tests* are concerned we hold that pattern (1) is adequate. Professor Kelley employs the elaborate pattern in the above table and interprets the common factor α as "heterogeneity, maturity, sex, and race." We argue that the factors α, β, γ, δ, ϵ, ζ, etc. are insignificant in these four tests, and that whatever common factor is found may be regarded as g. (Italics added.)

Professor Kelley (*loc. cit.*) charges me with the following errors:

(*a*) A false presentation of his argument because other elements of the problem were not considered (p. 365).

(*b*) Unawareness that when more variables are added to a given set the pattern explanation becomes more involved (p. 365).

(*c*) Overlooking this point in the interpretation of CAVD (p. 365).

(*d*) Implied charge that I (or people like me) would take four tests with common g, then take four other tests with a common g^1 and from these two results say $g = g^1$ (p. 366).

---

[1] Holzinger, Karl J. Thorndike's CAVD is full of *G*. *Journal of Educational Psychology*, March, 1931.

With regard to point (*a*) I may say that I was aware I had taken only a part of Professor Kelley's data. Evidence of such awareness is in the paragraph cited above. I was not presenting any argument of his either falsely or otherwise.

I am aware and was aware of the fact that when more variables are added to a set, the factor pattern may become more complicated. Thus, if we have an explanation for four variables x_1, x_2, x_3, and x_4, this explanation may not suffice when we add x_5. Professor Kelley notes that we may get a *group* factor between x_1 and x_5. He does not, however, explain how the addition of more variables to the original tests with no group factor, introduces several group factors *in the original form.* An explanation is possible, and would have been more to the point than the artificial data in his Table I, p. 365.

As regards the CAVD test I see no justification for the charge that I was unaware that other tests might have given a different result. Professor Kelley's remark on p. 366 is particularly misleading. "To generalize from the relationships found in four tests to relationships supposed to lie in the mind of man begs the question, for doing so involves the assumption that the four tests sample the entire mental life." I agree with this quotation as a statement of a general principle, but not if it is applied as a criticism of what I have done with the CAVD test and three others (which were all Professor Thorndike furnished me). Note that if the number "nine" be substituted in place of the number "four" in this quotation, the same criticism applies in less degree to Professor Kelley's own data. Surely the more tests we have and the more nearly they cover the areas of mental life, the better the opportunity for a complete picture of such life.

My own study showed that CAVD is full to the extent $r = .96$ with the same thing that three other so-called intelligence tests contain to a large extent. On p. 96 of his measurement of intelligence Professor Thorndike makes the following inference from the same correlations I employed. "Intellect CAVD is very much the same thing as that which is measured by representative examinations for so-called intelligence." All I have done is to show the same fact more precisely and to indicate how much CAVD is saturated with this "same thing" present in other tests of so-called intelligence.

The charge that someone might say $g = g^1$ when these two are obtained from different tests is too absent to more than recognize as such.

I may say in conclusion that I am much more interested in getting at the truth about factor analysis than in getting the better of an argument. An argument, however, may bring to light the crucial aspects of a theory a little more vividly than one sided presentation. The point here emphasized I believe to be, that factor patterns are functions of the tests we use They are also functions of the groups we test and of many other things. Except for this point I shouldn't have written this reply because I am certain Professor Kelley and I are in substantial agreement on all points at apparent issue

THE INFLUENCE OF JAZZ AND DIRGE MUSIC UPON SPEED AND ACCURACY OF TYPING

MILTON B. JENSEN

Western Kentucky State Teachers College

This study was made at the Training School of the Central State Teachers College, Michigan, three weeks before the close of the spring semester, 1930. Fifth, eleventh and twelfth graders (twelve boys and thirty-eight girls) in three typing classes numbering seventeen, seventeen and sixteen respectively were used as subjects All had had thirty-seven consecutive weeks of typing instruction in the high school In age they ranged from fifteen to twenty-two years (average, 17 88). Speed and accuracy were measured under three conditions, hereafter called states of distraction:

1. *Normal*—This was relatively free from distractions, being typical of the regular class procedure Passing through the corridors was infrequent during these periods and there were few of the noises common to schools in many of our population centers Whether this latter fact means that our subjects were more susceptible to distraction than they would be under city conditions we cannot say

2 *During the Playing of Jazz Music on an Edison Phonograph, Cabinet Model.*—A heavy steel needle was used in playing all the records, the tempo was kept normal and no changes were made in volume control. The jazz selections were: "Valencia," Perfect Record #14625, played by the Mayflower Serenaders, "At the Prom," Victor Record #38105A, played by Irving Mills and his Modernists, and "Bugel Call Blues," Victor Record #38105B, played by Jack Pettis and his Pets.

3. *During the Playing of Dirge Music*—The selections reproduced were· "Death of Ase," Victor Record #35470B, played by the Victor Concert Orchestra, "Thais Meditation," Victor Record #6186A, played by Fritz Kreisler accompanied by Carl Lamson, and "Indian Lament," Victor Record #6186B, played by Fritz Kreisler accompanied by Vincent O'Brien.

Practice effects were controlled by testing each of the three classes separately under all three conditions on three successive days and varying the order of testing so that, when the data for the three classes were combined, equal amounts of practice would accrue to each method Thus class "A" was tested in the order: Tuesday—normal, jazz, dirge, Wednesday—jazz, dirge, normal, and Thursday—dirge, normal, jazz Class "B" was tested in the order: Tuesday—jazz, dirge, normal; Wednesday—dirge, normal, jazz; and Thursday—normal, jazz, dirge. And class "C" was tested in the order: Tuesday—dirge, normal, jazz, Wednesday—normal, jazz, dirge; and Thursday—jazz, dirge, normal. Since the classes were approximately equal in size and since the testing all came at the same hour of the day, this procedure may be thought of as supplying an adequate control of practice effects

Three separate five minute typing tests were used, all being administered in the same order to each of the classes each of the three days: Test No. 4, p. 64, "New Rational Typewriting" by R. P. Sorelle, published by the Gregg Publishing Co , tests put out by the Royal Typewriting Co. in May, 1929 and April, 1930. Estimates of the reliabilities of these tests were secured by correlating the scores of the students in words per minute and applying the technique devised by Dr Shen in 1924 [1] These reliabilities in the order, Sorelle, Royal 1929, Royal 1930, were: under normal conditions, 62, 74, .57 (av., .64); during the playing of jazz, .89, .55, 68 (av , .71); during the playing of dirge, 70, 85, .46 (av., 67).

It will be seen that these examinations are accurate enough to justify their use as measures of group differences Assuming the reliability of each test equal to the average of the three tests, for each of the three states of distraction, an application of the Spearman-Brown formula gives a reliability of .84 for an averaging of the test scores under normal conditions, a reliability of .88 during the playing of jazz, and of .86 during the playing of dirge music [2]

While the differences in these estimates of reliability are of no great statistical importance they *do* show that the accuracy of our measures is not impaired by the distractions used What change there is, is in the direction of improved accuracy—an increase of 04 with jazz over the normal and of .02 for dirge over the normal.

Since typing results, traditionally, are measured in words per minute the effects of the distractions upon the reliabilities of the tests were determined in this unit. In determining the effects upon student performance, however, we have examined the speed of striking the keys (number of strokes) and the number of errors separately. Table I gives the average number of strokes, errors and words per minute for each of the three states of distraction, together with the probable errors of these means.

---

[1] Shen, E The Reliability Coefficient of Personal Ratings *Journal of Educational Psychology*, Vol. XVI, 1925, pp 232–236.

[2] Kelley, T L, "Statistical Method," p 205

TABLE I.—AVERAGE NUMBER OF STROKES, ERRORS AND WORDS PER MINUTE FOR TYPING UNDER THREE STATES OF DISTRACTION

	Normal	Jazz	Dirge
Strokes	215.08 ± 3.70	215.80 ± 3.82	209.99 ± 3.28
Errors . .	.937 ± .040	1.212 ± .082	.907 ± .053
Words .	33.64 ± .92	31.06 ± 1.04	32.03 ± .86

It is seen by inspection that jazz music had no appreciable influence upon the typing speed of our subjects as measured in strokes per minute. The dirge, however, brought about an average decrease from the normal of approximately five strokes per minute—2.37 per cent. This mean difference (5.09) divided by its standard error (2.41) gives 2.11. In terms of normal probability areas, the chances are nine hundred eighty-three in one thousand that the dirge as a distraction decreased the speed of striking the keys. When the number of errors per minute of typing is examined, however, a different picture is presented. The dirge, which materially reduced the speed of striking the keys, actually reduced the number of errors, leaving the end result (number of words per minute) little different from the normal. Jazz, as a distraction, markedly increased the number of errors—.56 per cent as against .44 per cent for the normal. The mean increase in errors per minute (.275) divided by its standard error (.038) gives 7.24, leaving no doubt as to the seriousness of the influence of jazz music on typing, so far as errors are concerned. The mean number of words per minute is not appreciably lower with dirge as a distraction than under normal conditions, but jazz, despite its lack of effect upon the number of strokes per minute, seriously affects the final score in words per minute, because it is conducive to error. Our subjects, on the average, typed 2.58 fewer words per minute with jazz than under normal conditions—a decrease of 7.67 per cent. The standard error of this difference is .796 so that the difference (2.58) divided by its standard error is 3.24.

In examining the relationship between speed and accuracy, the number of strokes and the number of errors were correlated (product-moment) for the three states of distraction. They were found to be: under normal conditions, − .111 ± .094; with jazz, − .191 ± .092; with dirge, −.508 ± .080 Had our populations been larger, more extensive examination would have been made of the ranges of ability most seriously influenced by distractions of this sort The relation between speed of typing and susceptibility to distraction is a problem of considerable importance to both the business man and the educator Unfortunately, our data cannot solve the problem as the writer should wish

It is of interest that speed and error are so little related under normal conditions A plotting of the percentages of error given in Table II shows a marked rise throughout the middle deciles as opposed to a gradual dropping off which one might be led to expect. This situation is even more marked with jazz as a distraction. With the dirge the tendency is still present but the decrease comes earlier—in keeping with the strong negative correlation between speed and error under this type of distraction. Table II gives the mean number of strokes per minute and the mean percentages of error for the deciles of the distribution under each of the three states of distraction. The decile order is the same throughout as under normal conditions, no changes being made for changed positions with jazz and dirge as distractions

TABLE II —MEAN STROKES PER MINUTE AND PERCENTAGES OF ERROR BY DECILES FOR THREE STATES OF DISTRACTION

Decile	Normal		Jazz		Dirge	
	Strokes per minute	Per cent error	Strokes per minute	Per cent error	Strokes per minute	Per cent error
1	149.8	45	148.2	59	56.0	54
2	179.9	42	192.4	44	181.8	46
3	152.0	40	200.3	78	186.7	30
4	202.2	55	196.7	77	196.4	49
5	208.2	58	208.5	.60	204.3	69
6	215.8	50	215.6	59	208.7	52
7	224.2	68	222.2	.97	213.4	58
8	231.8	30	234.2	39	228.7	32
9	255.6	17	247.9	26	243.8	35
10	281.4	21	285.8	35	272.1	27

CONCLUSIONS

Music is a serious distraction to typists under the conditions employed in this study. Jazz and dirge were used because they are extremes. Other types of music might be expected to have effects intermediate between the influences of these two It seems reasonable that any performance calling for rapidity and skill of movement will meet interference from stimuli as intruding as the extremes of jazz and dirge employed in this study.[1] The magnitude of this interference is indicated by our results. To what extent accommodation might offset the distraction over a protracted period of time, we have no measure The safe procedure, where speed and accuracy are demanded, is to avoid distractions of this type unless it can be demonstrated that accommodation has taken place to such an extent that performance will not be materially affected

---

[1] The writer has made a preliminary study of the effects of the radio upon simple arithmetic computations. Data for sixteen subjects adding single digit numbers during sixteen fifteen-minute experimental periods (eight with the radio and eight without—alternated to obviate practice effects and distributed over eight days) are strictly in keeping with the results obtained from the typing classes. Speed was not materially affected but there was a decrease of eight in the percentage of accuracy—ninety-three per cent of the additions attempted without the radio were correct and only eighty-five per cent of those attempted with the radio going were correct

A PROOF THAT THE POINT FROM WHICH THE SUM OF THE ABSOLUTE DEVIATIONS IS A MINIMUM IS THE MEDIAN

PAUL HORST

Personnel Research Department, United States Civil Service Commission

It is quite generally known that the median value of a statistical series is that value from which the sum of the absolute deviations is a minimum. Elementary handbooks on statistical method sometimes refer to this fact but since they are not concerned with derivations, they offer no mathematical proof. Both Yule[1] and Kelley[2] demonstrate the truth of this statement by showing that the absolute deviations would be greater if taken from any other point than the median

The writer rather incidentally discovered a proof which is more clean cut than any he has seen, and it might be of some interest to others Instead of showing that the sum of the absolute deviations would be greater if measured from any other point than the median we may show with a more general approach that the point from which the sum of the absolute deviations is a minimum is the median.

First we derive the general expression for the average deviation from any point. We have given a statistical series,

$$X_1, X_2, X_3, \qquad X_m, X_{m+1}, X_{m+2}, \ldots, X_n$$

arranged in order of magnitude, where all the values up through X_m lie below any point P, and all the values above X_m lie above this point. Then the average deviation about P will be,

$$\mathrm{AD}_P = \frac{1}{N}\left[\sum_{1}^{m}(P - X_i) - \sum_{m+1}^{n}(P - X_i)\right]$$

$$= \frac{1}{N}\left[mP - \sum_{1}^{m}X_i - (n - m)P + \sum_{m+1}^{n}X_i\right]$$

$$\mathrm{AD}_P = \frac{1}{N}\left[\sum_{m+1}^{n}X_i - \sum_{1}^{m}X_i - (2m - n)P\right] \qquad (1)$$

which is the same as formula (8) given by Kelley.[3]

---

[1] Yule, G. U. "Introduction to the Theory of Statistics" P 144

[2] Kelley, T L. "Statistical Method" P 74

[3] Kelley, T L. "Statistical Method" P 73

Now we wish to determine P in (1) so that AD_P will be a minimum To do this we simply differentiate (1) with respect to P and equate the derivative to zero. Thus,

$$\frac{d(AD_P)}{dP} = \frac{-(2m - n)}{N} = 0 \qquad (2)$$

Then solving for m in terms of n we have simply

$$m = \tfrac{1}{2}n \qquad (3)$$

That is, half the values are below P and half above, which is, of course, the definition of the median. Hence that point from which the sum of the absolute deviations is a minimum is the median

RELATION BETWEEN USE OF DIFFERENT PARTS OF SPEECH IN WRITTEN COMPOSITION AND MENTAL ABILITY[1]

CHARLES P LOOMIS AND ANNA MAY MORAN

North Carolina State College of Agriculture and Engineering

Since speech is used directly or indirectly more than any other single human attribute as a criterion by which to determine individual intelligence, it would appear advantageous to ascertain the relation of speech to mental ability from every possible angle. Most intelligence tests at the present time demand either speech reactions or reactions to speech stimuli, or both The problem undertaken in the present study is to discover the relationship between the proportion of different parts of speech in writing vocabularies of individuals and their mental ability

I

The importance of language in human activity is stressed by most psychologists, sociologists, and cultural anthropologists It is generally considered to be the chief vehicle by which the experience of the past and present is transmitted. Speech has made human culture possible and man a "time binding animal" To quote from Professor Cooley: "Language is the vehicle by which human relations exist and develop—all the symbols of mind, together with the means of conveying through space and preserving them in time, expression of fact, attitude and gesture, tones of voice, printing, and everything in the way of mental growth, has its existence therein Without language the mind does not develop a true human nature, but remains in an abnormal and non-descript state neither human nor properly brutal."[2]

So closely is language related to intelligence and the processes of developing mental ability, that it is believed that if one could measure an individual's language ability he would have a fair index to that individual's general mental ability. "It has often been said that thought would be impossible without words, and it is true that we can hardly conceive of human thought save as formed and embodied and expressed in language. Thought and articulate speech grew up, so to say, side by side; each implies the other, they are two sides of the same phase of mental development."[1]

---

[1] The article is a summary of a master's thesis presented at the N C State College, under the direction of Dr Karl C. Garrison, Professor of Educational Psychology.

[2] Cooley, Charles Horton: "Social Organization" Charles Scribner's Sons, York, 1909, p 62

A great many studies have been made to determine the relation between vocabulary and speech ability and general intelligence. For instance, Terman found correlations ranging from .85 to .95 between his vocabulary test scores and the scores made on the Stanford Revision of the Binet test.[2] Lockwood[3] found the correlation between composition scores based upon the Hudelson scale and the semester grades of freshmen to be .77 ± .029. Lockwood also found a correlation between general intelligence (Otis test) and composition scores of fifty-seven boys to be .67 ± .05, and of forty-three girls to be .76 ± .04, and concluded that a high degree of intelligence is directly related to ability to write adequately. Sangren[4] arrived at a similar conclusion after a study of ninth grade pupils; but Boss,[5] who studied twelfth grade pupils of superior mental ability, reported a lack of correlation between intelligence and achievement in English composition.

Closer to the problem at hand is the analysis and classification of different parts of speech used at different periods of language growth.

---

[1] Titchener, E. B. "A Beginner's Psychology." The Macmillan Company, New York, 1917.

[2] Terman, Lewis M. *et al.* "Genetic Studies of Genius," Volume I. Stanford University Press, 1925, p. 25.

[3] Lockwood, H. R.: Correlation of Mental Maturity of One Hundred College Freshmen and Their Ability to Write English Composition. *Unpublished Master's thesis,* Department of Education, University of Chicago, 1925.

Among studies of this kind are those of Hughes, W. H. The Relation of Intelligence to Vocabulary and Language Training. *English Journal,* Vol. XIV, Oct., 1925, and Buller, A. D. Intelligence Tests and Achievement in English Composition. *Educational Administration and Supervision,* Vol. XIII, Jan., 1927, and Garrison, K. C. The Relationship between Three Different Vocabulary Abilities. *Journal of Educational Research,* Vol. XXI, Number 1, Jan., 1930, pp. 43–44.

[4] Sangren, Paul V.: Intelligence Tests and the Classification of Students in Ninth-grade English. *Educational Administration and Supervision,* Vol. IX, Dec., 1923, p. 547.

[5] Boss, Mabel E.: The Relation of Performance in Mental Tests to Achievement in High School English. *Unpublished Master's thesis,* Department of Education, University of Minnesota, 1925, p. 89.

In this regard, discrepancies in the classification of certain portions of language activity have arisen because of the difficulty of assigning the part of speech to the same word or expression under different conditions. A summary of the literature on the relative number of the various parts of speech was made by C. W. Waddle, in 1918, with the following findings.[1]

1. Interjectional speech was characteristic at the beginning.
2. Nouns were used early in relatively large numbers.
3. From the first year on, the verbal element was relatively large.
4. The proportion of adjectives to adverbs was greater at younger ages.
5. Personal pronouns, relative pronouns, and subordinating and connecting words, were acquired with difficulty.

These studies, however, have been confined to the investigation of the *growth* of the use of the different parts of speech. Boyd,[2] in his study on the development of sentence structure in childhood, devotes a section to the elements in the vocabulary. He says: "Coming now to the words which are components of clauses and sentences, we may consider the relative percentages of the different parts of speech in the total words of the collected sentences of the child (counting each word each time it occurs)."

The percentages of the different parts of speech of the child studied by Boyd do not differ greatly from those of the study which follows, but they are classified differently. His comparison of the various parts of speech used by the growing child with those of male and female novelists is interesting. The percentage of relational words (conjunctions and prepositions) in adult sentences are quite higher, but the percentage of verbs and adverbs associated with them is considerably lower. Nouns and pronouns together are of about the same frequency in children and adults. Adults tend to use more nouns than pronouns, but the children use more verbs than the novelists. There was not much difference between the men and women novelists with regard to percentage of parts of speech used.

In a recent monograph, Symonds and Lee present data gathered for the purpose of analyzing the development of vocabulary usage in written composition, and to determine the changes that occur in the use of words as the child becomes more and more mature. In their conclusion they say: "It was surprising that the per cent of changes are so slight. A few basic articles, conjunctions, prepositions, pronouns, nouns, and verbs seem to be the framework of the language and are used whenever language is used."[1] One might anticipate from this conclusion that there would be little variation in the proportions of different parts of speech used by different ages and intelligence This was found to be somewhat true in the present study, but there was enough variation about the averages to warrant further statistical treatment involving correlations.

---

[1] Waddle, C. W.: "An Introduction to Child Psychology." Houghton-Mifflin Company, Boston, 1918, p. 71.

[2] Boyd, William: *British Journal of Psychology*, Vol. XVII, 1927, pp. 181–191.

In a study of high school pupils' vocabularies, O'Brien[2] found that the number of adjectives used by students increased steadily as the school grade increased. This was also found to be true of words or phrases pertaining to abstract qualities or ideas. He states that "it was felt that the extent to which the pupil used well chosen adjectives would provide a significant index of the precision and clearness of his thinking as well as a gauge of that phase of his vocabulary." Of course it would be hard to treat the "well chosen adjectives" quantitatively, but the increase in the number of adjectives used as higher school grades are attained might indicate that there is a relation between mental ability and proportion of adjectives used.

II

The data for the present study were obtained from the Junior division of the New Raleigh High School, Raleigh, North Carolina. Composition work of three groups, two eighth and one high seventh grade, was used. The composition papers were collected at different intervals throughout the first term of the school year 1929–1930. In order to allow the students to have as much freedom as possible in the use of their vocabularies, the subject-matter for the compositions was left to the writers. As a consequence, the compositions vary widely in their content as a statement of their analysis show, on p. 469

In order to keep as close to the aim of spontaneity as possible, it was necessary to reject some themes which had obviously been written by pupils who had resorted to reference books for facts in order to write the particular theme handed in. Long lists of proper or common nouns were also omitted.

---

[1] Symonds, P M. and Baldwin Lee. Studies in Learning of Expression, Number III—Vocabulary. *Teachers College Record,* Vol XXXI, October, 1929, p. 50.

[2] O'Brien, F P.: The Vocabulary of High School Pupils in Written Composition *Journal of Educational Research,* Vol II, 1925, pp. 334–350.

As a mode of procedure the following classifications of speech were set up: (1) Nouns and pronouns; (2) verbs and verb forms; (3) adjectives and adverbs (modifiers); (4) conjunctions and prepositions (connectors); and (5) the articles *a*, *an*, and *the*. In order to make the division of the words into different categories as objective as possible, words were assigned to their division only after ascertaining how they were used in the sentence. Sidwell and Siegfried's Handbook of Grammar[1] was used as a guide.

TABLE I.—ARBITRARY ANALYSIS OF THREE HUNDRED TWENTY-SIX COMPOSITIONS WRITTEN BY EIGHTY HIGH SCHOOL STUDENTS WHO WERE ALLOWED TO CHOOSE THEIR OWN FORM AND SUBJECT-MATTER

GENERAL THEME OF COMPOSITION	NUMBER SECURED
Description—nature, topics of interest, experiences, animal life, persons, holidays, geographical localities	107
Civics—tariff, Pan-American congress, League of Nations, construction of the Government, duty as citizens, current events of interest	91
Narrative—completion of stories, conversations, Story of Evangeline	64
Exposition—"How to Make," etc., "How to Earn Money," "How to Live at Home," "Why Study Latin"	27
Letters—personal, and letters of application	20
Moral Essay—interpretation of stories with moral	17
Total	326
Average number of compositions per pupil	4.075

After the words of each student's composition had been classified, the percentages of the total number of words used by an individual classified under the different catagories of the parts of speech were found, for each individual separately. In all cases all words were classified so that the sum of the percentages of the different parts of speech used by each individual always amounted to one hundred. The table on p. 470 shows the averages and deviations of individual percentages of different parts of speech used. Articles have the highest coefficient of variability.

In order that relationships could be found between uses of different parts of speech and intelligence, certain indices of the mental abilities of the students whose compositions were analyzed had to be obtained. The *first* measure used was the Terman Group Test of Intelligence, Form A. Mental ages were computed and used throughout the study.

---

[1] Sidwell, Paul and Russell G. Siegfried: "Handbook of Grammar." Charles Scribner's Sons, New York, 1928.

TABLE II.—AVERAGE AND VARIABILITY OF PERCENTAGES OF WORDS WHICH WERE CLASSIFIED UNDER FIVE DIVISIONS OF SPEECH AS TAKEN FROM THREE HUNDRED TWENTY-SIX COMPOSITIONS WRITTEN BY EIGHTY HIGH SCHOOL STUDENTS

	Arithmetic mean of per cents used by each person	Standard deviation	Coefficient of variability[1]
Nouns	35.2	2.095	5.95
Verbs	23.8	2.624	11.00
Modifiers	14.6	2.366	16.12
Connectors	16.7	2.003	11.96
Articles	9.5	2.290	24.01

[1] Coefficient of variability was found by the formula $CV = \frac{100SD}{mean}$.

The *second* measure used was a "Sentence Vocabulary Test," devised by Garrison of State College, Raleigh, and designed to cover a grade range of from third to eighth.[1] For a *third* criterion, the average grades in all thought subjects (English, history, civics, a natural science, and language) made by pupils during the first term of the school year 1929 and 1930 were obtained from the file in the principal's

TABLE III.—AVERAGES AND VARIABILITY OF MEASURES OF MENTAL ABILITY AND FATHER'S OCCUPATION OF EIGHTY HIGH SCHOOL STUDENTS

Measures of intelligence	Arithmetic means	Standard deviation	Coefficient variability
Mental age	13.75	1.545	11.236
Vocabulary test	47.05	6.134	12.702
Average school grades	83.10	7.807	9.395
Father's occupation	13.08	2.690	20.565

---

[1] This test is much like the Holley Vocabulary Scale, but made more difficult. An example of the test is as follows: To *pledge* is to . sedge cushion council promise

A study involving its use may be found in Garrison, K. C., The Relationship between Three Different Vocabulary Abilities. *Journal of Educational Research*, Jan., 1930.

office Also for an index of home environment, the father's occupation was used and given a quantitative rating using the Barr Scale Rating of Occupational Status[1] A table of these indices on p 470.

It was desirable to see what part the factor of sex played in the use of different parts of speech as well as the intelligence ratings. The following table shows that there is, on the whole, very little sex difference in the use of the different parts of speech

TABLE IV —COMPARISON OF THE PERCENTAGES OF DIFFERENT PARTS OF SPEECH USED BY EIGHTY HIGH SCHOOL GIRLS AND BOYS

	Number	Nouns	Verbs	Modifiers	Connectors	Articles
Boys	39	34.0	21.0	10.1	16.7	14.3
Girls	41	35.5	23.7	15.0	16.8	9.0

It may be seen by the above table that the chief difference is the boys' and girls' use of speech in modifiers and articles The girls use 4.9 per cent more modifiers than do the boys, while the girls use 5.3 per cent fewer articles than do the boys

TABLE V —COMPARISON OF MENTAL AGES, VOCABULARY TEST SCORES, AND AVERAGE SCHOOL GRADE BY SEX

	Number	Mental age	Vocabulary ability	Average school grades
Boys	39	13.78	47.5	80.6
Girls	41	13.57	48.3	85.4

Table V compares the sexes by three other criteria. The mental age and vocabulary index of the boys and girls show very little difference. The girls outranked the boys by 4.8 per cent in their average school grades

---

[1] In this list one hundred representative occupations, each definitely and concretely described, were rated from one to one hundred by thirty judges according to the grade of intelligence which each was believed to demand Probable error values were computed for all these one hundred occupations and the ratings distributed. In the present study those occupations that were not included in the one hundred listed in the Barr Scale were classified as the occupations most like them The Barr Scale Ratings list is fully described in Terman's *Genetic Studies of Genius*, Vol. I, p. 66

By way of further analysis, mental age, vocabulary tests, average school grades and the percentages of the different parts of speech were studied and zero order coefficients were computed to show relationships among these factors. Table VI shows the various relationships found by this computation.

TABLE VI —THE RELATION BETWEEN THE DIFFERENT PARTS OF SPEECH AND MENTAL AGE, VOCABULARY TEST SCORES, AND AVERAGE SCHOOL GRADES

Parts of speech	Mental age	Vocabulary test score	Average school grades
Articles[1]	52 ± 06	48 ± 06	51 ± 06
Connectors	20 ± 07	17 ± 07	14 ± .08
Verbs	− 27 ± 07	− 17 ± 07	− 22 ± 07
Modifiers	− 10 ± 08	− 05 ± 08	08 ± 08
Nouns	.09 ± 08	−.10 ± 08	− 06 ± 08

---

[1] A correlation coefficient of 30 ± .07 was found to exist between English Grades and the per cent of articles used

So far as the sample used is concerned, Table VI shows:

1 The use of the articles *a*, *an* and *the*, is a better index to mental ability than is the use of any other part of speech.

2 The use of connectors next to articles shows the highest positive correlation with measures of mental ability, but the coefficients are not high.

3. The use of many verbs indicates lower mental ability with low coefficient.

4. Other parts of speech do not indicate much relationship with the criteria of mental ability used in the study.

TABLE VII —INTERCORRELATIONS BETWEEN THE DIFFERENT PARTS OF SPEECH USED IN WRITTEN COMPOSITION BY EIGHTY HIGH SCHOOL STUDENTS

Parts of speech	Articles	Connectors	Verbs	Modifiers	Nouns
Articles		38 ± 07	− 18 ± 08	46 ± 06	58 ± .05
Connectors			− 49 ± 06	− 22 ± 07	−.04 ± 08
Verbs				− 36 ± 07	01 ± 08
Modifiers					− 36 ± 07

In the table above may be found a study of the interrelations among the usage of different parts of speech.

The study of the intercorrelations among the different parts of speech summarized in Table VII indicates the following facts concerning compositions studied:

1 The use of a great many words of any single part of speech, resulted in a less or nearly the same number of words of another part of speech (with the exception of articles).

2. The use of articles correlates positively with the use of modifiers, nouns, and connectors, but negatively with verbs.

In order that the relationship between the children's environment, as measured by occupation, and other attributes might be studied, various correlation coefficients between measures of mental ability and the usage of the parts of speech were computed. Data are presented in Table VIII.

TABLE VIII—RELATIONSHIP BETWEEN THE USE OF ARTICLES, CONNECTORS, MODIFIERS, AND THE MENTAL ABILITY, VOCABULARY TEST SCORES, AVERAGE SCHOOL GRADES AND FATHER'S OCCUPATION

Articles	Connectors	Modifiers	Mental age	Vocabulary test	Average school grades	Father's occupation
Articles	38 ± 07	.10 ± 07	52 ± 00	48 ± 06	51 ± 00	32 ± 07
Connectors		22 ± 07	20 ± 07	17 ± 07	11 ± 08	02 ± 08
Modifiers			− 10 ± 07	− 05 ± 08	08 ± 08	05 ± 08
Mental age				58 ± 05	27 ± 07	21 ± 07
Vocabulary test					40 ± 00	30 ± 00
Ave school grades						01 ± 08
Father's occupation						

The data presented in Table VIII corroborates many other studies that have found high mental ages associated with high vocabulary test scores and school grades. The table also indicates the following

1 Father's Occupation as rated by the Barr Scale Ratings of Occupational Status[1] previously referred to, correlates positively with the use of articles, which is the best measure of intelligence so far as parts of speech are concerned

2 Father's Occupational Ratings correlate positively and significantly with mental age and vocabulary test scores As here rated, father's occupation does not seem to influence school grades

---

[1] See footnote 1, p 471

III

By way of conclusion it may be stated that:

1. It is difficult to explain the correlation between the measures of mental ability and the use of articles. The correlation coefficient between the percentage of articles used and mental age was .52 ± .06, between the percentage of articles and the vocabulary test score, .48 ± .06; and between the percentage of articles and average school grades, .51 ± .06. Since there is not the chance of using articles incorrectly that there is with most other parts of speech, such as general modifiers, the percentage of articles used makes a better index of mental ability. This is especially true in view of the fact that the correlation coefficients between measures of mental ability and the use of articles are in most cases higher than correlation coefficients between the different measures of mental ability themselves. The only exception to this is the correlation between vocabulary test scores and mental ages, .58 ± .05, while the coefficient between mental age and the use of articles is .52 ± .06.

2. The positive correlations between the use of connectors and mental ability would be easier to explain than that of articles with mental ability, if one assumed that the first part of a sentence had to be kept in mind while the last part was being composed. The correlation coefficient between percentage of connectors and mental age was .20 ± .07, between percentage of connectors and vocabulary test scores, .17 ± .07, and between percentage of connectors and average school grades, .14 ± .08.

3. As would be expected in any structure, the parts of which are as interdependent and constant as the parts of a sentence, an increase in the use of one part of speech will automatically be accompanied by a decrease in the use of the other parts of speech. This holds true for all parts of speech except articles. Since articles are so closely related with nouns and modifiers, positive correlations would be expected between articles and these parts of speech, and such relationships actually exist.

4. The use of articles has almost as much relationship with occupational status as has vocabulary ability as represented by the vocabulary test scores—the former relation being represented by the correlation coefficient .30 ± .06, and the latter, .32 ± .07. Children's environment and possibly heredity, as represented by the occupational ratings seems to influence the mental ability as represented by the mental age and vocabulary test scores, and at the same time their

use of articles. School grades are not significantly influenced by parent's occupation in this study.

5 A thoroughly quantitative method of measuring the correct and incorrect usages of different parts of speech, rather than the study of percentages of words classified as being certain parts of speech, would probably throw new light on the subject of the relation between use of different parts of speech and mental ability. It is hoped that such a method can be developed and applied.

FALSIFICATION OF AGE: A FACTOR IN CHILD GUIDANCE[1]

ANNA COHEN AND NATHAN ALTROWITZ

Psychiatric Social Workers, Child Guidance Department, Hebrew Sheltering Guardian Society, Pleasantville, N. Y.

The writers have recently had the good fortune to cooperate in the handling of a case involving a factor in child guidance which is often obscured; it is that of falsification of age. This apparently harmless act may result in innumerable conflicts which give rise to abnormalities in the development of a personality.

Our instance is that of a boy who was committed to the institution as unmanageable. He is an illegitimate child whose mother is considered neurotic. The boy was born in a distant city, and the putative father deserted almost at once. After moving to her present residence, the mother married a man who is a drunkard. The mother repeatedly told the boy that he must help her and his younger brother; that he was really her mainstay because of the incapacity and worthlessness of the stepfather. The boy was jealous of his brother, yet he took a serious big-brotherly interest in him. He quarreled with his stepfather, was disobedient, and wilful. He was also known to steal frequently. When he arrived at the institution, he continued to react in much the same manner for a long time. In order that he might sooner take over the responsibilities of the home, his mother had made him believe that he was two years older than she thought he was. She explained her placing him at the institution at the lower age by saying that only by doing this could she get the institution to care for him longer and so make it possible for her to provide at least for her younger son and herself.

Thus the boy was officially considered as old as his mother supposed him to be. That even this official age was incorrect will be seen later, though whether the mother herself did not know or would not reveal his correct age cannot be determined.

His maladjustment at the institution resulted in his being placed under the care of the Child Guidance Department. The workers noticed that there was a worried and somewhat resentful look on his face, his step was heavy and his head sagged At school, he was considered a poor student, doing little work, and coming tardy frequently. But the social aspect of the problem was even more serious He stole, lied, quarreled, was disobedient and did not usually mingle or play with the other children. He exhibited great glee at his success in appropriating the possessions of others without being caught, and if he were found out, he did not seem to mind.

---

[1] The writers wish to express their appreciation to Dr. Leon W. Goldrich, Executive Director; Dr. Samuel Z. Orgel, Psychiatrist, Miss Julia Goldman, Head Psychiatric Worker; and Mr. Myron B. Blanchard, Psychiatric Social Worker, all of the institution referred to in this paper, for their kind suggestions concerning the article.

Since there was a conflict between the figures for the age as his mother had impressed it on him and the age as found in the institution's records, the workers communicated with the authorities at his birthplace. It was learned that he was even a year younger than his mother supposed, that is, three years younger than he had come to believe himself to be

The boy was informed of his correct age, and gradually a marked change was noticed. He said that he no longer felt physically inferior to other boys as formerly. His entire mein has become more carefree; his head is held higher, his step is lighter, and he runs and plays with the others. He also feels that he is not retarded in school, since he is in the proper grade for his age His attitude toward school has changed; he makes it a point never to be tardy, takes a keen interest in his school work and is having no difficulty with it. Improvement is also noticeable in his social adjustment. He no longer steals, nor is he disobedient, and he seldom quarrels

This description indicates that here at least is one cause of problems that can be dealt with directly, quickly, and adequately. It emphasizes the significance of the apparently trivial matter of giving an incorrect age to attain some special end without taking into consideration the fact that its total effect on the personality may be very far reaching.

Yet how often, for example, do the parents, misleading the authorities, start the child in school before he has reached the proper age? The causes of falsification of a child's age vary Those who are responsible for such an act are generally motivated by one or more of several definite factors

The mother may have no time to care for the child, she may have to go out to earn a living, or she may have a younger child to take care of. On the other hand, the child may have been unwanted He may be an illegitimate child, whose illegitimacy will not be suspected after the age has been falsified Once this is done, there develop problems that are difficult of solution.

First of all, the child is forced into situations that he is too young to meet Imagine the discouraging and asocializing effect of constant defeat in athletics in competition with much bigger boys of his purported age. It halts the child's play life and tends to make him seclusive

Second, it requires that the child live up to two divergent standards of behavior, one at home and one at school. So there develops in his mind a conflict as to which one he is expected to uphold.

Third, it raises the child from one social status to another without adequate preparation. At the time when he should be looking forward to completing his education, he is placed in the position of breadwinner. The child therefore feels deep concern over his failure to gain the training necessary to make him capable of shouldering these responsibilities

The combined effect of all these factors, impressed upon the growing child may eventually turn him into a serious problem.

The solution of the difficulty must necessarily lie in the return of the child into the family situation at his correct age with the standards suitable for that age.

NOTE —Since the purpose of this paper is to present the problem of age falsification *per se*, no attempt has been made to give the analytical interpretations that would be involved in completing the picture of the case

NEW PUBLICATIONS IN EDUCATIONAL PSYCHOLOGY AND RELATED FIELDS OF EDUCATION

CONDUCTED BY FRANCES M FOSTER

Adolescent Education, by Frederick E. Bolton. New York: The Macmillan Company, 1931. Pp. 506.

It is perhaps no exaggeration to say that there is no other field of education concerning which more scientific knowledge is needed, and less is available, than in the field of adolescent education. There has been a dearth of scientific literature produced in this field during the past fifteen years

"Adolescent Education" by Professor Frederick E Bolton will meet a long felt need in this particular field. The educational world apparently has not fully recovered from the loss sustained by the departure of that remarkable student, teacher and educator to whose memory Professor Bolton has dedicated this book.

The author states that the book is intended to be an analysis and inventory of the adolescent for the purpose of finding his potentialities and his needs as determined by his unfoldment The course of study is treated only incidentally and only as a means to an end The author tells us that business, social and economic conditions and traditions have been the educational determinants, and that the educational theorist, not trained in psychological analysis proceeds with what he conceives to be the logical unfoldment of the subjects and arranges the traditional school subjects into what he conceives to be ideal "mental discipline"

The child psychologist, instead of securing his basic information from industries and the world at large, endeavors to secure his information directly from the children This procedure consists, perhaps first of all, of a thorough scientific analysis of the child's physical equipment, accompanied by a thorough and scientific analysis of his instinctive tendencies, his mental equipment, and his interests all of which serve as a key to the kinds of activity best suited to the proper development of the child

Professor Bolton points out the often unobserved fact that a scientific educational procedure builds upon what *is*, in order to achieve what is to be. This necessitates a scientific and functional knowledge of the pre-adolescent, adolescent and the post-adolescent life of the child Professor Bolton's book consists of sixteen chapters each giving a scientific treatment of a fundamental aspect of adolescent education. The book reveals a mastery of the scientific literature in its field Each chapter is accompanied by a carefully selected bibliography. The book concludes with a thorough and enlightening treatment of the ever present and gigantic problem of Character Education.

DENNIS C. TROTH.

Pennsylvania State College.

A CORRECTION

Editor, THE JOURNAL OF EDUCATIONAL PSYCHOLOGY.
Dear Sir:

There is an error (of transcription) in my formula in the article, "The Sigmas of Combined Distributions Calculated from Sigmas, Means, and Frequencies of Component Distributions," in the April issue of THE JOURNAL. In Formula 7, page 310, the second term of the left-hand member representing the combined mean should be squared This error was discovered by Doctor Louis W. Gellermann. I am also indebted to Doctor Harry Helson for reporting difficulty with the formula as it was printed.

The correct form of the formula is·

$$\frac{(\sigma_1^2 + \bar{x}_1^2)N_1 + (\sigma_2^2 + \bar{x}_2^2)N_2 + \ldots (\sigma_n^2 + \bar{x}_n^2)N_n}{N_1 + N_2 + \ldots N_n} - \left(\frac{(\bar{x}_1)N_1 + (\bar{x}_2)N_2 + \ldots (\bar{x}_n)N_n}{N_1 + N_2 + \ldots N_n}\right)^2 = \sigma^2_{1,2,\ldots n} \quad (7)$$

C. R GARVEY.

YALE UNIVERSITY

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII October, 1931 Number 7

ONE MORE STUDY OF PERMANENCE OF INTEREST*

HARVEY C. LEHMAN

Ohio University

AND

PAUL A. WITTY

Northwestern University

The question of permanence of vocational interest is obviously of utmost importance to those concerned with problems of personnel and guidance. If guidance be based upon a pupil's assumed occupational interest, and if this assumed interest proves subsequently to have been merely a passing fancy or a temporary whim, the outcome of course will be wasteful and the practice injudicious. Obviously the question of the degree of permanence of interest is one of great individual and social significance.

In order to ascertain the extent and degree of permanence of vocational interest, the writers administered the Lehman Vocational Attitude Quiz to a large number of school children in Topeka, Kansas, and in Kansas City, Missouri.†

The Vocational Attitude Quiz consists of a comprehensive and catholic list of two hundred occupations. First, the children are asked to check *only* those occupations in which they are willing to engage as life work. They are then asked to indicate: (1) The three occupations which they would like best to follow, (2) the one occupation which they most likely will follow, (3) the three occupations which they think are the best money-makers, (4) the three occupations which they believe are most respected, and (5) the three occupations which they believe will require the least amount of effort. The number of children from whom data were obtained is indicated in Table I.

---

* The present study presents some findings that have been made possible by a grant-in-aid from the Social Science Research Council.

† The writers are indebted to Miss Anna G. Myers, Mr. J. F. Kaho, Superintendent A. J. Stout, and Superintendent George A. Melcher for assistance in securing these data.

TABLE I.—TOTAL NUMBER OF CHILDREN INCLUDED IN A SERIES OF INVESTIGATIONS OF CHILDREN'S VOCATIONAL ATTITUDES

Ages	Topeka, October, 1927		Topeka, May, 1928		Kansas City, November, 1928 (superior economic status)		Kansas City, November, 1928 (inferior economic status)		Kansas City, November, 1928 (average economic status)	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
8½	203	269	82	101	137	180	140	174	304	324
9½	340	358	257	274	212	211	182	203	342	369
10½	403	372	347	340	261	238	236	222	430	435
11½	387	375	385	355	259	272	253	228	356	377
12½	302	372	420	337	331	326	260	260	313	332
13½	305	346	363	378	373	317	235	222	252	298
14½	301	301	352	320	291	345	166	164	246	282
15½	324	322	285	343	204	305	132	181	209	241
16½	251	290	239	201	248	217	93	95	172	197
17½	162	150	160	171	132	111	40	37	108	95
18½	57	36	69	44	48	33	13	7	43	19
Totals	3254	3286	2968	2933	2580	2555	1765	1799	2773	2959

Total boys	13,340
Total girls	13,532
Grand total	26,878

Table II presents the order of merit (rank frequency) of occupations in which boys 8½ to 18½ years of age stated that they *would be willing to engage*. Table III presents the rank of occupations *liked best or preferred most* by the boys 8½ to 18½ years of age.

Table II is to be read as follows: "Aviator" is the one occupation that was most frequently checked by boys of 11½ to 18½ years of age as the occupation in which they would be willing to engage as a life work. This activity ranked sixth for boys 8½, third for the boys 9½ years of age, and second for the boys of 10½ years of age. A blank space in the table indicates that the particular occupation was mentioned by an exceedingly small percentage of pupils, and a rank therefore was not assigned Tables II and III show the rank only of the most frequently mentioned occupations for each age group

TABLE II —RANK IN FREQUENCY OF OCCUPATIONS IN WHICH BOYS OF 8½ TO 18½ YEARS OF AGE WOULD BE WILLING TO ENGAGE

No	Occupations	Ages										
		8½	9½	10½	11½	12½	13½	14½	15½	16½	17½	18½
66	Aviator	0	4	2	1	1	1	1	1	1	1	1
81	Civil engineer		28	22	18	10	5	3	1	2	4	2
18	Architect				38	27	16	7	4	3	3	3
83	Electrician or electrical engineer			37	28	18	12	0	3	4	2	4
100	Cowboy	1	1	1	2	2	2	11	33	35		
36	Sailor	2	5	3	5	7	15	21	29	38		
33	Army officer	3	2	0	4	5	8	17	10	15	18	13
34	Soldier	4	4	4	9	11	22	30				
110	Sheriff or policeman	5	7	10	12	23	31					
70	Professional baseball player	7	0	5	3	3	1	5	11	11	12	20
07	Jockey or automobile racer	8	8	9	7	8	13	14	18	10	25	27
92	Ship builder	0	11	10	23	33						
74	Fireman or engineer on a train	10	9	0	10	13	11	10	10	20	31	
102	Fisherman, hunter, or trapper	11	11	11	13	15	18	15	23	26	20	
70	Fireman (answering fire alarms)	12	12	19	23	31						
0	Banker	11	13	11	17	17	11	12	7	6	7	8
35	Naval officer	13	10	7	0	6	6	9	8	10	11	16
3	Lawyer	15	18	21	21	21	20	18	12	8	5	5
00	Professional boxer or wrestler	16	16	16	15	20	21	25	31	30		
08	Physical director or athletic coach			42	33	20	27	20	20	17	18	6
52	Musician	30	43	33	27	20	23	19	19	24	17	8
65	Radio expert	20	24	23	22	19	10	10	0	11	8	0
26	Automobile dealer	21	17	26	21	28	30	28	25	21	18	10
103	Explorer		31	20	19	16	17	13	13	0	0	0
107	Forest ranger or woodsman	42	25	18	14	12	7	4	0	5	6	11
108	Detective or secret service work	31	22	12	8	1	3	2	5	7	7	15
01	Stockraiser or ranchman	25	14	11	11	11	9	8	14	13	20	10
80	Chemist or chemical engineer					32	24	20	17	16	13	10
10	Buyer for a large store						28	33	21	12	15	7
136	Confectioner	18	33	11								
109	Night watchman	17	20	13								

Those of equal frequency are given the same rank but, contrary to the usual custom, ranks are omitted in instances in which few or no boys checked the occupation

Certain of the frequency ranks change markedly with advance in chronological age. For example, in Table II the third item from the top of the list is "Architect." This occupation was rarely checked by young children, but at ages 16½ to 18½, it was checked frequently and therefore received rank three

In Table III, the second item is "Cowboy" This occupation receives rank one at ages 8½ and 9½. At the upper age levels of course this occupation ranks very low. The findings for this occupation are portrayed clearly in Fig 1 and Table IV which present the percentage of five groups of white boys of various ages who asserted that they would be willing to become cowboys. It will be noted from Fig. 1 and Table IV that the ranks are relatively consistent at the several early age levels. Nevertheless, it is clear also that permanence of interest in this particular vocation is decidedly lacking.

TABLE III—RANK IN FREQUENCY OF OCCUPATIONS LIKED BEST BY BOYS OF 8½ TO 18½ YEARS OF AGE

No	Occupations	Ages										
		8½	9½	10½	11½	12½	13½	14½	15½	16½	17½	18½
00	Aviator	2	2	1	1	1	1	1	1	1	1	1
100	Cowboy	1	1	2	2	2	5	10				
3	Lawyer	0	7	10	6	7	0	10	5	1	2	3
81	Civil engineer		10	8	0	3	2	2	2	3	3	7
2	Doctor (physician, surgeon or specialist)	11	13	10	11	11	11	11	7	6	0	0
33	Army officer	4	3	3	4	0	15	24	12	21	13	8
70	Professional baseball player	0	0	5	3	4	1	5	0	10	12	11
0	Banker	7	0	12	15	11	13	13	0	7	10	10
53	Musician	20	15	14	10	10	12	7	8	0	6	5
35	Naval officer	12	10	0	5	0	13	14	15	18	15	10
83	Electrician or electrical engineer			20	22	5	3	3	3	2	4	4
18	Architect					17	10	1	4	5	5	2
08	Physical director or athletic coach						21	23	12	11	11	4
41	Movie actor	17	11	18	18	10	16	10	13	10	10	13
74	Fireman or engineer on a train	8	7	7	7	0	7	0	17	12		
34	Soldier	3	4	4	12	20	30					
110	Sheriff or policeman	10	11	15								
30	Sailor	5	5	0	8	15	21	20				
07	Jockey or automobile racer	13	8	0	13	13	17					
108	Detective or secret service work			13	10	8	0	8	11	14	15	
14	Newspaper work							25	18	14	14	7
107	Forest ranger or woodsman			20	10	13	8	0	10	8	17	13
05	Radio expert					23	23	15	19	10	21	11
00	Farmer, miscellaneous		10	17	20	19	18	10				
01	Stockraiser or ranchman			10	14	12	11	12	14	11		
60	Professional boxer or wrestler			19	17	22	25	30				
84	Mechanical engineer										7	
80	Chemist or chemical engineer										8	

The writers have assembled, similarly, data for all the occupations of the check-list which were reported by the children as ones which they would be willing to enter At age 8½ the average number checked by the boys was 23.3; at age 18½, the average was 15.6. Since the number of occupations in which boys may engage is exceedingly large, the percentages for each of the many occupations are small. Because of this fact, the data showing percentages for each age level do not portray vividly the striking changes in attitudes that really take place from year to year. A careful study of the data suggests however that permanence of interest does not exist in many or even most of the vocational preferences of boys Indeed, from examination of the ranks of the occupations at different age levels, the writers have been forced to conclude that perhaps no phase of

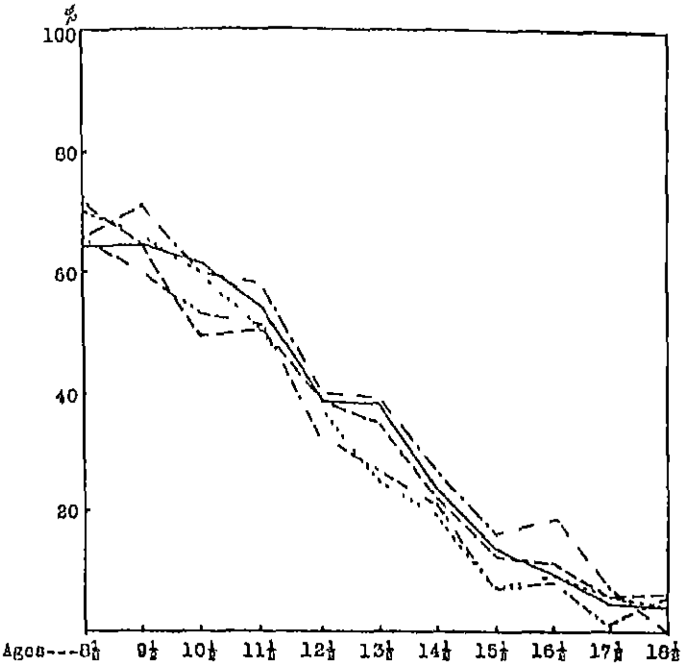


Fig 1.—Per cents of five groups of white boys of various age levels who expressed willingness to become "Cowboys"

———————— Topeka, Oct , 1927
———————— Topeka, May, 1928
———————— Kansas City, Mo , Nov , 1928 (Superior economic status)
———————— Kansas City, Mo , Nov , 1928 (Inferior economic status)
———————— Kansas City, Mo , Nov., 1928 (Average economic status)

human nature is subject to such marked change as that reflected in the vocational interests and preferences of growing boys. This conclusion differs strikingly from that expressed by numerous workers of the past and by several present-day ones.

Among the earliest studies of permanence of interest were those of Thorndike.[1] Thorndike concluded that it would be hard to find "any feature of a human being which was a much more permanent fact of his nature than his relative degrees of interest in different lines of thought and action." (P. 456.)[1] He concluded also that interest and ability are so closely bound together that "either may be used as a symptom of the other almost as well as for itself." (P. 456.)[1]

TABLE IV.—PERCENTAGE OF FIVE GROUPS OF WHITE BOYS OF VARIOUS AGE LEVELS WHO EXPRESSED WILLINGNESS TO BECOME "COWBOYS"

Ages	Topeka, October, 1927	Topeka, May, 1928	Kansas City, November, 1928 (superior economic status)	Kansas City, November, 1928 (inferior economic status)	Kansas City, November, 1928 (average economic status)
8½	64	71	66	66	70
9½	64	64	59	71	66
10½	61	49	53	60	60
11½	54	50	51	58	50
12½	38	40	33	40	37
13½	38	35	27	30	25
14½	24	22	21	26	19
15½	14	13	7	16	7
16½	10	12	8	18	10
17½	5	6	1	7	6
18½	4	7	6	0	5

Several writers interpreted Thorndike's conclusions as implying that children's vocational ambitions are relatively permanent and that vocational ambition is a suitable index of vocational ability. In 1922, Freyd wrote regarding Thorndike's work: "The methods used were such as to make results of limited significance yet in their field these results are practically all that are available." (P. 244.)[2] Several writers seem to regard Thorndike's former pronouncements and the subsequent generalizations of other writers as valid when they are applied to children today. Moreover, there are many attempts to verify the basic theses. Books and articles are now being published which present evidence of a continuance of interests for a few months or a single year as proof that such alleged vocational interests are symptomatic of permanent interests Franklin reports data obtained from some fifteen hundred junior high school pupils of Baltimore who were asked in December, 1922 to name the occupation that they would like best to enter Follow-up questionnaires were filled out in May, 1923, October, 1923, and December, 1923. The following conclusions were expressed

> The vocational interests of junior high school pupils show a very high degree of permanence over a period of at least one year,—and that a very critical one Two children out of every three have the same preference at the end of the period as they had at the beginning (P 155)[3]

McCracken and Lamb report a series of studies that were made in Denver March 1, 1920, May 25, 1920, and in April, 1921. The problem of permanence of interest was studied for a period *less than three months in length.*

> Of 4481 junior and senior high-school students whose choices were recorded for March 1, 1920, and May 25, 1920, with the intervening time nearly three months, 73.9 per cent of the boys and 79 5 per cent of the girls chose the same occupation on both occasions [4]

The credulous reader is informed that "These choices, although made so close together, do indicate some permanency " (P 43.)[4]

It will be of interest at this point to examine Fig. 1. If the lines of this graph represent the situation as it will exist from year to year it may be asserted that a very large percentage of the boys of a given age who are willing to become cowboys will continue to cling to this notion three months hence,-- or even a year hence. To this extent, these data corroborate those of Franklin, and those of McCracken and Lamb. It is clear, however, that *over a longer period of time,* a permanence of interest would not be found It therefore seems invalid to base conclusions regarding permanence of vocational interest upon a study that continues *only for a few months or for a year*

The writers have attacked the problem of permanence of interest in a somewhat different manner They have studied the vocational interests of several thousand children of ages 8½ to 18½. The writers are here assuming that the present 8½-year-old children would respond next year as do the 9½-year-old ones, etc The validity of this assumption they are at the present time investigating Several follow-up studies are under way to test the reliability of the children's reports. Nevertheless, it seems logical to assume that groups of 8½-year-old children will respond next year as do children now 9½ years old provided that the children live in a conservative city, not affected appreciably by a shifting population. If the written expression of willingness (and intention) to enter a given vocation be regarded as expression of interest, it is obvious that many of the interests of the children herein studied cannot be regarded as permanent since many interests of the children change markedly during a period of several years.

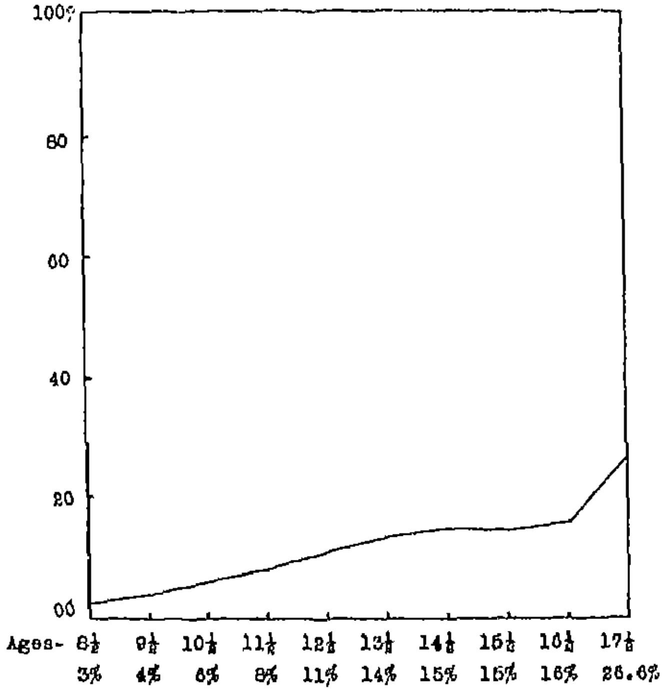


Fig. 2.—Per cents of five groups of boys who stated that they expected to enter one of the following types of engineering: chemical, civil, electrical, mechanical, or mining

It will be of interest to examine the hypothesis that interests are symptomatic of ability. Ability is sometimes conceived as inherent capacity; inborn and unchangeable. If this concept be adhered to, the hypothesis that interests are symptomatic of ability postulates the additional hypothesis that interests do not change markedly. The data herein presented reveal that certain of the children's interests change markedly during a period of several years If ability be conceived as inborn and predetermined, it is clear that the children's interests can not be considered reliable indices of their abilities

Figure 2 shows the percentage of the five groups of boys who stated that *they expected to enter* one of the following forms of engineering. Chemical, civil, electrical, mechanical, and mining. The United States census report shows that in 1920 these five groups of engineers included only 135,789 white persons. According to the census report there were, in the United States in 1920, 29,053,677 white males of ten years of age and above who were gainfully employed. Of this number the five engineering groups listed above comprised less than one-half of one per cent (0 16 per cent) If the ratio between the number of engineers and the total number of other gainfully employed white males continues to remain similar to the 1920 condition, it is clear that no age group can contribute many members to the five engineering groups previously mentioned. Figure 2 shows that 26.6 per cent of the 17½-year-old boys state that *they expect to enter engineering.* Of course one must bear in mind that many of the less competent boys have been eliminated from school by the time age 17½ is reached It is also well to bear in mind that these are city boys having unusual educational advantages. After due allowance has been made for such facts as these, one can scarcely conceive of the demand for engineers increasing sufficiently to absorb so many youthful aspirants

The Vocational Attitude Quiz included forty professional endeavors. Figure 3 presents the percentage of the five groups of boys who stated that they *expected to enter* these forty professions The percentage varies from 24 to 50 at the various ages. According to the U. S. census report, these forty professions included in 1920 only 3.15 per cent of the total number of white males of age ten and above who were gainfully employed. There is of course no valid reason for believing that the 1920 ratio is to exist indefinitely. On the other hand, it is impossible to believe that over fifty per cent of the 14½-year-old boys will have the opportunity to enter the professions It is more likely that only a small proportion of these boys will be able to find a place in the professions and a still smaller proportion will enter upon the particular profession which at age 11½ they expect to enter.

The situation for the girls of various age levels who stated that they *expected to become stenographers or typists* is similar to that of the boys. Figure 4 shows that over thirty per cent of the 18½-year-old girls expressed this vocational desire. The U. S. census report shows that in 1920 the white female stenographers and typists (562,664) included only 9.3 per cent of the white female workers (6,062,216) of ages ten and above.

As a pioneer in the field of vocational and educational guidance, Thorndike is entitled to the respect of all. Nevertheless, Thorndike would insist, the writers believe, that his data and conclusions be verified. Certainly, in a country in which social attitudes and values are subject to such reactionary and sporadic fluctuation, it is essential that behavior data be assembled repeatedly to show important trends and socially significant changes. It is, however, difficult for us to modify attitudes which have been generally accepted. Striking indeed is the domination of outgrown, impotent, and socially undesirable determiners of conduct.

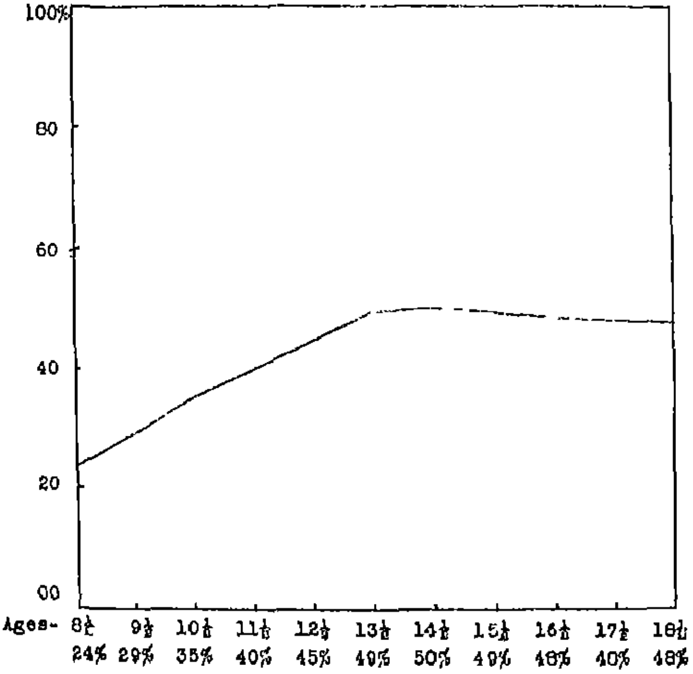


FIG. 3.—Per cents of five groups of boys who stated that they expected to enter one of forty professions

The obstacles to psychological advance include not only factors such as inadequate methodology, social taboo, etc. Psychology is held back also by a serious factor which is not frequently mentioned. Once an error has crept into psychological literature, it often requires the labor of several generations to modify or to dislodge it. At some future date it will be of interest to students of psychology to trace numerous generally accepted present-day misconceptions and rationalizations. At that time it is highly probable that many of the studies of permanence of vocational interest will be cited as splendid examples of the inadequacy of the retrospective method for identifying interest.

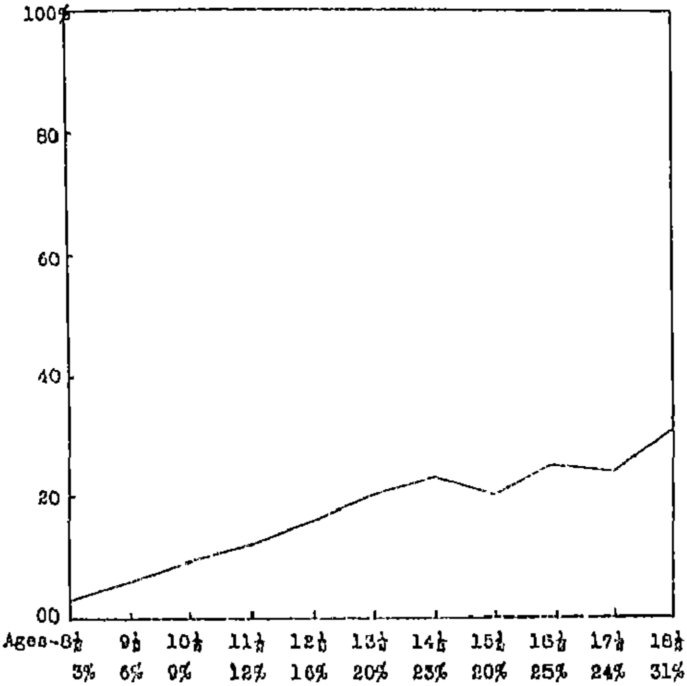


Fig. 4.—Per cents of five groups of girls who stated that they expected to become stenographers or typists.

The data assembled by the writers indicate that certain vocational interests cannot be permanent. (See Figs. 2, 3, and 4.) Other interests may be more permanent. And available data show quite clearly that interests are symptomatic of ability only to a limited degree.

REFERENCES

1. Thorndike, E L · The Permanence of Interests and Their Relation to Abilities *Popular Science Monthly*, 1912, Vol LXXXI, 449–456.
2. Freyd, Max: A Method for the Study of Vocational Interests. *Journal of Applied Psychology*, 1922, Vol. VI, 243–251.
3 Franklin, E E · The Permanence of the Vocational Interests of Junior High School Pupils *Vocational Guidance Magazine*, 1927, Vol V, 152–156.
4 McCracken, T C and Lamb, H. E "Occupational Information in the Elementary School" Boston· Houghton Mifflin Co , 1923, pp. xiv–250

MIRROR READING AS A METHOD OF ANALYZING FACTORS INVOLVED IN WORD PERCEPTION*

MILES A TINKER AND FLORENCE L GOODENOUGH

University of Minnesota

This study was undertaken for the purpose of ascertaining to what extent the nature and frequency of the errors made in learning to read aloud English prose material in the reversed position as shown in a mirror might throw light upon the perceptual processes of ordinary reading Four adult subjects, one man and three women, none of whom had had previous formal experience with mirror reading, participated in the experiment. The subjects worked in pairs, one reading aloud while the other acted as recorder and timekeeper A rectangular mirror eight by ten inches was used. In the initial stages of the experiment the mirror was held in an upright position by the recorder; but as this was found to be somewhat inconvenient for both persons, a support was devised which held the mirror at a constant angle of one-hundred degrees with the top of the table. Each reader held the book himself and adjusted its distance from the mirror according to his own preference.

In taking the record, the recorder occupied a position facing the reader who held the book with its back towards himself and the pages facing both the mirror and the recorder. In this way the subject was able to read from the mirror, while the recorder followed the text directly and made note of each error as it occurred A prepared form was used for this purpose. At the end of ten minutes, which was timed with a stop watch, the signal to stop was given and a record was made of the total number of words read and the number of errors made This included errors which were spontaneously corrected by the subject as well as those which passed unnoticed or at least without correction.

Reading was carried on for ten minutes daily during the period from April 4-July 31. At this time the summer vacation intervened, and practice was suspended until September 25 when it was again resumed and continued until November 1 Save for occasional unavoidable omissions, practice took place seven days a week during these periods

---

* Grateful acknowledgment is due to Dr Mary Shirley who served as subject throughout the experiment, and to Mrs Eva Tinker who not only acted as subject but is also responsible for the work of tabulating the results

There was some variation in the time of day at which the reading was done, but as a rule it was carried out during the early part of the evening.

The reading material used during the spring and summer included two detective stories and one novel.* During the fall, approximately half of "Psychology, Its Facts and Principles," by H. L. Hollingworth, was read. The last type of material was better suited to our purpose than were the earlier selections, because of its greater uniformity from page to page. It is probable that of the irregularity in the character of the fiction material used for the first part of the experiment, with its frequent changes from conversation to descriptive matter, accounts at least in part for the irregularity of progress in the early portion of the learning.

Marked individual differences existed among the four subjects in the ability to adapt themselves to the reversed image in mirror reading, and these differences increased rather than decreased with practice. All four subjects showed rapid improvement during the early stages of the learning and very slow gain thereafter There is a distinct lapse in efficiency upon taking up each new book. This difference is to be attributed chiefly to the large number of new words which were encountered within the early stages of reading a new book, rather than to a change in the typography The size and style of type was very similar for the four books

For the entire experimental period the mean number of words read per error was as follows: For Subject No 1, 138, Subject No. 2, 347, Subject No. 3, 153; and Subject No. 4, 141. With the exception of Subject No. 2, there is but small difference in accuracy among the four readers

While the foregoing results have some interest for the general subject of learning, the main interest has to do with the nature, frequency and position of the errors made, and the relationship of these

---

* In order of reading, the books used were:

1. Gluck, S "The Green Blot."
2. Jordan, E. G.: "The Blue Circle"
3. Vance, L. J., "White Fire"
4. Hollingworth, H L.: "Psychology Its Facts and Principles"

The first and third were read completely by all four subjects Subjects No 2 and No. 3 read about half of the second book, then substituted "The Closed Book" by J. E Meyer. Approximately half of this was read Subjects No 1 and No 4 completed the second book without substitution factors to word-perception. A summary of the experimental findings to date* reveals some lack of agreement as to the relative importance for perception of total word-form and of the separate elements or letters of which a word is composed, or in other words, whether words are commonly apprehended as configurational wholes or by successive perception of their separate parts or letters. The consensus of opinion, however, seems to be in favor of the former hypothesis The experimental finding of Cattell,[1] Messmer,[6] Goldscheider and Müller,[3] Erdmann and Dodge,[2] Pillsbury,[7] and others are in conformity with this viewpoint

Although it seems fairly well demonstrated that practiced readers tend to perceive words as units, rather than as aggregates of smaller elements, the importance of the analytic method in building up these larger perceptual wholes must not be overlooked. The tendency to revert to a method of analysis in terms of single letters or syllables upon encountering a new or difficult word was very evident in all four subjects This appears to have been the major factor in producing the characteristic drop in the curve at the beginning of each new set of material where a large number of new words were suddenly introduced into the reading vocabulary. The changes were marked, not so much by increase in the total number of errors made, but rather by frequent hesitations upon encountering new words. At these times there was usually silent or audible spelling of the entire word or some part of it, or attempts at reading by syllables Examples follow:

Subject No. 4 upon encountering the word *lapse* on the seventh day of practice first hesitated while the word was examined, then spelled *l-a-p*, then, unable to decide whether the following letter was an *a* or an *s*, passed on to the final letter which was also pronounced aloud, *e* This made it possible to bridge the gap, and the word was correctly pronounced

That this procedure was frequently employed by all four subjects is evident both from their introspective reports and from such overt behavior as the oral spelling of words or syllables It was found to occur at two levels, *viz.*: in the apprehension of single words, as in the example given, and in the reading of sentences or phrases, when an entire word might be passed over temporarily to be later supplied by the context In both instances, however, it seemed probable that the context was not the sole factor in supplying the missing element, but rather that it functioned in connection with certain vaguely perceived elements in the total form of word or letter which were not at first sufficiently clear to permit unaided recognition. This viewpoint is in general agreement with the findings of Zeitler,[9] Winch,[8] and Korte.[5]

---

* M A. Tinker: Visual Apprehension and Perception in Reading. *Psychology Bulletin*, Vol XXVI, 1929, pp 223–240.

It is clear that the method just described may give rise to errors as well as to successes. This is illustrated by the treatment given to the word *roused* by the same subject on the same day. The first three letters *r-o-u* were spelled aloud as before, and again the *s* proved a stumbling block. In this case, however, perception of the final letters resulted in an incorrect interpolation of the missing part and the word was finally pronounced *rounded*. The extent to which this method has been employed may be inferred from the distribution of 1795 errors in which the general form of the substituted word was similar to that of the correct word, *i.e.*, differed from it by not more than two letters. For the four subjects combined, the number of such errors occurring at the beginning of the word was only one hundred sixty-four, as compared to seven hundred ninety-six at the end of the word and eight hundred thirty-five in some intervening position. The overwhelming importance of the initial portion of the word for total apprehension has been noted previously by Goldscheider and Müller,[3] Huey,[4] Pillsbury,[7] and others. The corroborative evidence offered by our study has special significance, since it shows that it is the fact of *initial position* rather than location at the right or left side of the word which is the determining factor. This factor is of sufficient strength to cause individuals whose previous reading habits have presumably taught them to give special weight to the letters at the left-hand side of the word, to reverse their habits completely with the reversal of the usual order of letters within the word. While the number of cases is too small to warrant generalization, it is interesting to note that the substituted word is somewhat less likely to involve a change in the context in the case of the two most rapid readers than with the two slower readers In thirty-one per cent of the errors of this kind made by Subject No. 1 and in twenty-five per cent of those made by Subject No. 2 the substituted word was an approximate synonym for the correct word, as compared to nineteen per cent and eighteen per cent for Subjects No 3 and No 4 respectively. This suggests the possibility that the more rapid readers tend to fill in their perceptual gaps upon the basis of larger context-units than do the slower readers. A somewhat similar situation in the apprehension of single words is suggested by the fact that the two most rapid readers make a greater percentage of the total number of errors in the middle of the word than do the two slower readers This may be the result of a somewhat greater tendency on the part of the more rapid reader to "jump over" larger areas in the apprehension of words, or, in other words, to identify words upon the basis of a smaller number of "determining" elements Further data with a larger number of subjects are needed to verify this hypothesis

Errors of inflection involving a change only in the prefix or suffix with the root word remaining unchanged, have been treated separately For the four subjects combined, there are eighteen instances of a change in the prefix as compared to three hundred thirty-three instances of a change in the suffix. The importance of the initial position is again clearly demonstrated

A total of one hundred forty cases in which one word was substituted for another of very different form but of similar meaning were made by the four subjects Of these ninety-four, or sixty-seven per cent of the total were made by Subject No 1 who was the most rapid reader. Again the use of larger context-units in the case of this subject is suggested.

Omission of letters or syllables follows the same general pattern as substitution of letters or syllables Out of a total of one hundred ninety errors of this kind, sixteen occurred at the beginning of a word, forty-four in the middle and one hundred thirty at the end Fifty-five per cent of the total number were made by Subject No. 1. There were one hundred forty-eight instances of omitted words, and fifty-six instances of inserted words One hundred forty of the former and forty-two of the latter were attributable to Subject No. 1.

Errors resulting from a reversal of the order of letters within the word, or within some part of the word were of fairly frequent occurrence. There were three hundred forty cases of complete reversal This category includes all cases in which the beginning and end of the word were interchanged, with or without accompanying reversal of intervening letters. Example: *was* for *saw*, *told* for *dolt*, *on* for *no*, etc. There were one hundred seventy cases in which the reversal affected only a part of the word. These have been classified according to the position of the reversed portion within the word In thirty-eight cases the order of the letters making up the first syllables was reversed on initial attempt or pronunciation, *e.g.*, *now* for the first syllable of *wondering*. Errors of this kind were likely to be corrected spontaneously, since as a rule their nature was such that the word could not be sensibly completed without correction. In one hundred sixteen cases the initial and final letters were correct, but a reversal was made within the word. *Form* for *from* and *vice versa* is a common example. There were sixteen cases in which the reversal occurred at the end of the word, as *until* for *unitl*. In thirty-four cases the order of words within a phrase or short sentence was reversed, as *is it* for *it is*, etc.

It is impossible to say to what extent these reversals are the direct result of interference from the normal reading-habit and to what extent they represent perceptual difficulties of a more fundamental nature. Every primary teacher knows the difficulty which most beginning readers experience in learning to distinguish between such words as *was* and *saw*, *fell* and *let;* on and *no*, etc. That interference factors were operative to some extent, at least, in the present experiment was evident both from the introspective reports of the subjects, and from their overt behavior. In the beginning of the experiment all subjects reported that the page not only appeared to be reversed from right to left but also to be upside down. This led to considerable difficulty in passing from one line to another or in going from page to page. One subject almost invariably began a new page at the bottom instead of the top, and for all there was a perceptible period of fumbling adjustment at the beginning of each new page during the early part of the experiment. All found it more or less difficult to shift from line to line; and not infrequently the device so often employed by little children under similar difficulties— that of following the line with the finger or of pointing to each successive line in order to keep the place was brought into use. In spite of these devices, and the early recognition of the existence of this particular type of difficulty by all four subjects, in twenty-one cases an entire line was omitted in reading; in seven cases a line was repeated, and in two additional cases the fixation point shifted to the adjacent line without awareness on the part of the subject, who, as a result, read half of one line and half of the following line.

The effect of interference is also seen in certain letter-confusions which were frequently made. It is a matter of common observation that normal children upon learning to read rarely have much trouble in learning to distinguish words or letters upon a vertical or up-and-down basis, but that the horizontal, or right-and-left distinction frequently causes much trouble. As Terman has expressed it ("Measurement of Intelligence," p. 177) "it is the "*p*'s" and "*q*'s" that children must be told to mind, not the "*p*'s" and "*b*'s " For all subjects, however, *d* and *p* were confused more frequently than any other letters, with *b* and *d* as a close second. The operation of the "up-side-down" illusion has almost certainly been the determining factor in the first instance Other letters frequently confused were *a* and *s*, *h* and *b*, and *w* and *m* In general it may be said that such confusion is most apparent in the case of letters made up of identical or closely similar elements, thus necessitating a distinction upon the basis of spatial position only

Although it is evident that a comparison of word-confusions or letter-confusions within material such as this, where individual words or letters occurred with unequal frequency, and therefore had unequal chances to be confused, does not permit too wide generalization, it has nevertheless seemed worth while to tabulate the word-confusions which were found to occur most frequently within the limits of our material The results of this tabulation are shown below

Of for *to* (or *vice versa*), 88, *on—no*, 68; *saw—was*, 60; *the—this*, 19; *that—this*, 14; *left—felt*, 12; *now—won*, 11; *tell—let*, 11; *the—that*, 9, *the—his*, 8; *face—fact*, 7; *physiological—psychological*, 7; *and—any*, 7; *though—through*, 6; *part—past*, 6; *him—his*, 5; *to—or*, 5; *ever—even*, 5, *whatever—whether*, 5, *and—but*, 3, *started—stared*, 3; *for—of*, 3, *put—but*, 3.

The findings of Erdmann and Dodge[2] and of Pillsbury[7] appear to show that word-length is an important factor in the determination of word-form. Zeitler[9] on the other hand, regards word-length as a factor of minor importance Our findings clearly bear out the former viewpoint. A tabulation of a total of 2812 errors (not including omission or insertion of entire words) yielded the following results.

No change in number of letters, 1505 cases, difference of one letter, 1027 cases; of two letters, 202 cases, of three letters, 55 cases, of four letters, 17 cases, of five letters, 3 cases, of six letters, 2 cases; of seven letters, 1 case

It will be noted that ninety per cent of the errors do not involve a change in word length of more than one letter An examination of the nature of the most frequent errors as reported in a foregoing paragraph will show that in none of these instances is the change in word length greater than one letter, while in eighty-five per cent of the errors included within this list, no change in word length has taken place.

Since only four subjects participated in the experiment, it was thought that it would be worth while to see to what extent individual idiosyncrasies might be affecting the general trend of the results, or, in other words, to see to what extent the distribution of the various kinds of error might vary from subject to subject We have therefore classified the errors made by each subject under the different heads discussed previously, and then computed the rank-order correlation between the comparative frequency of the various types of error for each subject with every other subject. The categories used in classifying the errors will be repeated here for convenience.

Omission of line
Repetition of line
Mixing up two lines
Omission of letters or syllables at beginning of word
Omission of letters or syllables at end of word
Omission of letters or syllables in middle of word
Substitution or addition of letters at beginning of word
Substitution or addition of letters at end of word
Substitution or addition of letters in middle of word
Reversal of entire word
Partial reversal at beginning of word
Partial reversal at end of word
Partial reversal in middle of word
Reversal of words or phrases
Phrases substituted or added

The inter-correlation between the order of frequency of the various types of error for the four subjects ranged from +.57 to +.85 with an average intercorrelation of +.75 In a similar manner we have computed the rank-order correlations between the frequency of various single-letter confusions for the four subjects, using only the twelve confusions most frequently made, and in this case differentiating between the two alternatives in each pair. The list of confusions follows: *b* for *d*, *b* for *p*, *d* for *b*, *d* for *p*, *f* for *t*, *p* for *b*, *p* for *d*, *a* for *s*, *h* for *b*, *w* for *m*, *t* for *f*, *s* for *a* In this case the inter-correlations ranged from +.26—+.78 with an average inter-correlation of +.59. This is in close agreement with the correspondence between individuals found in most of the reported studies on the relative legibility of individual letters.

SUMMARY

1 The results of a learning experiment in the reading of English prose from the reversed image shown in a mirror have been utilized for the study of certain factors influencing word-perception

2. Four adult readers participated in the experiment. Practice for ten minutes daily was carried out by each subject over a period of approximately five months. Marked individual differences in speed of reading were apparent from the beginning, and these differences tended to increase with practice

3 An analysis of the errors made in terms of their position within the word yielded results which are in conformity with those of certain previous investigators in the following respects (1) The initial letters were found to be of far greater importance in the apprehension of words than the final or the intermediate letters, (2) The context was frequently utilized in the final apprehension of difficult words or in deciding among the various alternatives suggested by imperfect or partial perception, (3) Word-length appeared to be a factor of major importance in determining word-form

4. Interference from the normal reading habit was shown, not only in the reversal of letters or of letter-elements in the horizontal direction which is a difficulty frequently found among children when first learning to read, but also in the presence of a definite "up-side-down" illusion which tended to produce similar confusion in the vertical direction

5. A fair degree of correspondence was found among the four subjects in regard to the relative frequency of the different kinds of error made The mean inter-correlation of the order of frequency of certain specified errors classified on the basis of their nature and their position within the word was +.75 for the four subjects; between the frequency rank-orders of the twelve most common letter confusions was + 59

REFERENCES

1. Cattell, J. M Ueber die Zeit der Erkennung und Benennung von Schriftzeichen, Bildern und Farben *Phil Stud*, Vol II, 1885, pp 634–650
2 Erdmann, B. und Dodge, R · "Psychologische Untersuchungen über das Lesen auf experimenteller Grundlage" Halle, 1898
3 Goldscheider, A, und Müller, R F Zur Physiologie und Pathologie des Lesens *Zsch F Klin Med*, Vol XXIII, 1893, pp 131–167.
4 Huey, E B, "The Psychology and Pedagogy of Reading" New York, 1909.

5 Korte, W.: Ueber die gestaltauffassung im indirekten Sehen *Zsch. f Psychol*, Vol XCIII, 1923, pp 17-82

6 Messmer, O: Zur Psychologie des Lesens bei Kindern und Erwachsenen. *Arch. f d ges Psychol*, Vol II, 1903, pp. 190-298.

7. Pillsbury, W. B: A Study in Apperception *American Journal of Psychology*, Vol VIII, 1897, pp. 315-393

8 Winch, W. H. Teaching Beginners to Read in England *Journal Educational Research Monograph*, No 8, 1925

9. Zeitler, J: Tachistoskopische Versuche ueber das Lesen. *Phil. Stud.*, Vol. XVI, 1900, pp. 380-463.

THE RETENTION OF MIRROR-READING ABILITY AFTER TWO YEARS

FLORENCE L. GOODENOUGH AND MILES A. TINKER

University of Minnesota

This study was designed to supplement an earlier investigation by the authors[1] on progress in learning to read ordinary prose material as viewed in a mirror

Four subjects participated in the original study. Practice was carried out for ten minutes daily over a period of approximately seventeen weeks.

Two years later, an experiment in relearning was undertaken. Ten minute practice periods were employed as before. Practice was continued for twenty consecutive days. The book used was "Psychology, Its Facts and Principles," by H. L. Hollingworth The first half of this book had been read at the end of the original learning study. For re-learning the second half was used, thus insuring similarity of subject-matter without the use of identical material. All four of the subjects participated in relearning None had had any intervening practice.

A comparison of the number of lines read per day during the re-learning period with the corresponding number read during the original learning is presented in Table I.

In this table the mean number of lines read during each successive five-day period of the relearning experiment is compared with the final stages of the original learning in which the first part of the same book was used. The number of days practice on this book was the same for both learning and relearning.

The rank-order of the four subjects is the same for both learning and relearning, if the average for the twenty days of practice is considered. With few exceptions, this is also true when the data are grouped by successive five-day stages Two of the four subjects make a distinctly higher score on the relearning than was made during the final period of learning, in spite of the fact that there had been no intervening practice. The remaining two subjects made approximately the same average scores on both learning and relearning.

---

[1] Tinker, M. A and Goodenough, F L. Mirror Reading as a Method of Analyzing Factors Involved in Word-perception. THIS JOURNAL 1931, Vol. XXII, pp 493

TABLE I.—COMPARISON OF LEARNING AND RELEARNING SCORES IN MIRROR READING

Subject	Days, 1–5		Days, 6–10		Days, 11–15		Days, 16–20		Average of all trials	
	L	*R*	*L*	*R*	*L*	*R*	*L*	*R*	*L*	*R*
1	131.8	130.4	139.4	142.2	115.6	130.0	150.4	138.2	141.8	142.2
2	80.2	81.2	82.6	102.8	80.8	99.6	82.4	93.0	81.5	94.1
3	75.2	80.0	78.6	90.0	82.2	88.6	81.1	87.6	79.4	88.8
4	65.2	74.0	74.8	81.4	81.8	77.8	94.4	80.8	79.1	78.5

L, mean number of lines read per day during the original learning period.

R, mean number read during relearning (different portions of the same book used).

One gained 0.4 of a line, the other lost 0.6 of a line. When the average score for the first five days of relearning is compared with the last five days of the original learning, all subjects show a slight loss, but this loss was more than recovered by the end of the second five-day period of relearning. These results indicate that with adult subjects, retention in mirror-reading tends to persist with practically no loss in efficiency for a period of at least two years.

VOCATIONAL INTERESTS AND TYPES OF ABILITY

RALPH H GUNDLACH AND ELIZABETH GERUM

University of Washington

This paper presents estimates of the inter-correlations between fifteen of the vocations measured by Strong's Vocational Interest Blank, and offers a partial analysis of the types of interest which are in part responsible for the degree of correlation.

We have several times found a need for a knowledge of the inter-relations between the various occupational groups measured by the Vocational Interest Blank. The manual of directions contains figures of inter-correlations for about twenty pairs of vocations; but there are stencils and norms for almost that many occupations The manual, further, gives no indication as to the number of cases upon which the correlations are based, or the PE's of the correlations It would be an enormous task to work out the actual inter-correlations with an adequate number of cases.

Our own efforts were directed toward obtaining, fairly easily, an estimate of the actual correlations Instead of using scores made by various individuals upon the tests, we used the stencils themselves; instead of using the fully weighted scores, we used only the sign of the score, and instead of using the total blank, we used only four of the sub-tests (1A, 1B, 7 and 8). A relative measure of the degree of correlation in interests between two vocations in each of these sub-tests was obtained by determining the number of agreements in sign in the "Like," "Indifferent" and "Dislike" columns for all items; and from this figure was subtracted the number of agreements which would be expected purely by chance When these data from all the vocations and sub-tests used were added and assembled in a table, the rough relations between the various occupations could be seen, but their significance in terms of the coefficient of correlation could not, of course, be determined. To make a conversion possible, these figures were correlated with actually determined correlations between vocations. The actual correlations were obtained from those published in the manual of directions and from Correlations we had determined. Strong does not state the number of cases upon which his correlations are based Ours are based only upon fifteen to forty cases When a value for any pair of interests occurred in both sources, the average was taken. There were finally thirty-four different correlations,

TABLE I—ESTIMATED INTER-CORRELATIONS BETWEEN VOCATIONAL INTERESTS

$PE_{est.}$ is .075

	Lawyer	Journalist	Engineering	Certified Public Accountant	Advertiser	Architect	Doctor	Psychologist	Chemist	Personnel Manager	Purchasing Agent	Real Estate	Teacher	Minister	Life Insurance
Lawyer		.99	27	59	73	39	63	39	41	13	03	59	06	09	34
Journalist	99		47	53	97	48	56	36	003	10	19	.60	05	03	22
Engineering	27	47		48	38	56	56	39	.78	35	60	.13	31	−.15	−17
Certified Public Accountant	59	53	48		33	51	32	25	18	31	23	35	35	05	17
Advertiser	73	97	38	33		35	23	28	05	15	33	71	−10	−10	30
Architect	39	48	56	51	35		57	45	61	23	50	07	05	−10	−07
Doctor	63	56	56	.32	23	57		45	58	−16	20	19	−02	11	08
Psychologist	39	36	39	25	28	45	45		65	25	−05	−35	48	54	−35
Chemist	41	003	78	18	05	61	58	65		33	31	−24	19	02	−53
Personnel Manager	13	10	35	31	15	23	−16	25	33		33	03	54	38	15
Purchasing Agent	03	19	60	23	33	50	20	−05	31	33		36	−11	−17	11
Real Estate	59	60	13	35	71	07	19	−35	−24	03	36		−29	−30	83
Teacher	06	05	31	35	−10	05	−02	48	19	54	−11	−29		84	04
Minister	09	03	−15	05	−10	−10	11	54	02	38	−17	−30	84		01
Life Insurance	34	22	−17	17	30	−07	08	−35	−53	15	11	83	04	01	
Average	564	555	496	465	461	460	430	374	334	312	286	268	239	125	113

ranging from −.10 to 88, which were plotted against the corresponding estimates. The correlation between the actual correlations of interests, and the estimates of agreement, based upon the signs of the weights, is .918 ± .0175 By means of the obtained regression equation each of the values in the table giving the relative closeness of agreement of the vocational interests was converted into its most probable correlational value. Table I gives these estimated intercorrelations. The $PE_{est.}$ is .075. As a check, the differences between the thirty-four actually determined and the corresponding predicted correlations were plotted: The PE of the distribution was .079

Even a casual inspection of the table shows many indications of group factors. Since the actual basis of the degree of correlation between any pair of vocations can be found in the stencils, an analysis of factors involved could readily be made providing there was a significant set of categories for classifying the various items. Following Thorndike's suggestion that there are perhaps three types of intelligence, and assuming a close relation between interest and ability, we first attempted a classification in terms of the categories, Abstract, Mechanical and Social. It soon appeared that these classes needed amplification and subdivision. After several trials, what appeared to be a fairly complete classification, with five main headings and a total of twelve sub-headings was constructed. Three of the main headings correspond to Thorndike's Abstract, Social and Mechanical intelligence. The remaining headings collect the rather miscellaneous sub-groups that fitted in no particular place Nine different individuals, representing five departments of instruction, acted as judges. The judges classified the items in entire independence of each other; and in view of their complaints as to the difficulty of making the classifications, the amount of agreement among them is surprising The judges omitted a number of items which they held were not classifiable into these types of categories. Hence only the first four sub-tests of the Blank were eventually used The judges were asked to indicate which items they felt belonged under the following headings:

(*A*) Social interests
  1. Desire to dominate or boss a group (army officer)
  2 Desire to be exhibited before a group (actor)
  3 Desire to mingle in a crowd (picnics).
(*B*) Intellectual interests
  4. Empirical (physics)
  5. Abstract (*mathematics*).

(*C*) Technical interests
  6 Clerical (bookkeeper)
  7 Mechanical (auto repair man)
  8 Technical-professional (dentist).
(*D*) Creative interests
  9 Artistic (artist)
  10 Practical (music teacher)
(*E*) Interest in physical skill
  11 Competitive (auto racer)
  12. Non-competitive (taking walks)

Those items which four or more of the nine judges agreed in classing under the same heading were considered as properly placed. All others were omitted.

It remained to compare each of the stencils for the various vocations with this analysis of the interests. Any item in the Interest Blank may be marked "Like," "Indifferent" or "Dislike." For any particular interest any of these alternatives may be weighted much or little, and with a plus or a minus score All "Indifferent" scores were ignored. Only fairly heavily weighted positive or negative scores (four or five, depending upon how many such items appeared in the particular vocation) were considered. Score was totaled by counting one each for positively weighted "Like" items and negatively weighted "Dislike" items; and by subtracting one each for negatively weighted "Like" items and positively weighted "Dislike" items Accordingly it was possible in scoring one vocation to have a count of zero, one or two (plus or minus) for every item Table II presents scores obtained in this fashion for fifteen vocations for the various types of interest, converted into per cent of perfect score The Minister's score, *e.g*, of fifty per cent under Creative artistic interests, means that that profession had a score of seventeen on these seventeen items

Considering the various vocations as wholes, it first appears that none of the vocations can be said to depend only upon one or another of the major groups of interests. A very important factor seems to be what members of the vocations dislike. Real Estate Salesmen, and to a lesser extent, Lawyers and Journalists, are characterized chiefly by their almost universal dislikes, while the Ministers, and to a lesser extent the Teachers and Personnel Managers, are characterized by omnovorous likings. The interesting characteristics of the individual vocations are too numerous and too readily observed to be noted

We may now turn to the relations between the groups of interest If Thorndike's suggestion that there are three types of intelligence

Table II—Analysis of the Types of Interest for Each Vocation (in Per Cent)

Class	Social			Intellectual		Technical			Creative		Skill	
Number of items	B twenty-four	E thir-teen	M six-teen	E twenty-six	A fif-teen	C twelve	M twenty-six	T six-teen	A seven-teen	P eight	Ph eight	N four
Lawyer	−42	0	−19	10	0	−25	−38	−53	−15	−38	−25	0
Journalist	−25	0	−3	−4	−30	−46	−38	−56	6	13	−25	−50
Engineering	−10	−42	−9	21	30	−21	35	−9	−9	−6	−13	0
Certified Public Accountant	−17	−12	−3	0	27	29	−4	13	0	13	−13	−13
Advertiser	−19	−4	−31	2	−13	−71	−33	−44	15	0	6	0
Architect	−35	−23	−16	−6	−3	−58	4	−9	41	44	−6	−25
Doctor	−42	−27	−16	17	7	−63	0	−9	−3	0	−13	0
Psychologist	−25	−8	−6	+36	33	−21	10	0	12	38	−25	0
Chemist	−6	−15	−13	17	27	−33	29	22	0	19	−13	25
Personnel Manager	29	0	19	27	0	0	19	9	0	0	0	13
Purchasing Agent	4	0	−7	−25	−30	13	4	13	−35	−19	19	0
Real Estate Salesman	4	0	−13	−36	−53	−50	−58	−28	−44	−38	0	0
Teacher	2	−4	9	33	40	21	−2	−9	21	25	0	13
Minister	13	12	−13	35	50	13	4	13	50	25	−6	0
Life Insurance Salesman	17	0	13	−19	−20	0	35	0	0	0	13	0

Social, B—boss, E—exhibitor, M—unobtrusive mingler.
Intellectual, E—empirical, A—abstract
Technical, C—clerical, M—mechanical, T—technical professional
Creative, A—artistic, P—practical.
Skill, Ph—physical competitive, N—non-competitive.

is true; and if, as has been fairly well established, there is a close relation between abilities and interests, then the correlations between the classes within each of our major headings should be high, and the correlation between any member of one group and any member of another group should be low. To facilitate such a comparison, rank-order correlations between the classes of interests were worked out (Table III). Fairly high correlations obtain between the two Intellectual and between the two Creative classes of interests. The various Social classes and Technical classes, however, do not have such high intercorrelations. In fact, the averages both of the correlations of Clerical interests with the three Social classes, and of Practical Crea-

TABLE III.—INTERCORRELATIONS BETWEEN THE VARIOUS CLASSES OF INTERESTS (RANK-DIFFERENCE METHOD)

	Social			Intellectual		Technical			Creative		Skill
	B	E	M	E	A	C	M	T	A	P	Ph
Social											
B		53	.50	02	22	01	40	54	− 10	− 19	.69
E .	53		28	− 12	− 20	33	− 22	03	04	− 19	35
M. . .	50	28		10	− 07	60	36	30	05	16	20
Intellectual											
E .	02	− 12	10		89	26	34	.03	47	46	− 35
A	22	− 20	− 07	89		43	30	41	46	55	− 33
Technical											
C. . ..	01	33	69	26	43		28	58	00	13	19
M .	40	− 22	.36	34	30	28		03	11	25	.17
T..	54	03	36	03	41	58	03		09	32	25
Creative											
A .	− 10	04	05	47	46	09	.11	09		88	01
P.	− 19	− 19	.16	46	55	13	25	32	88		−.18
Skill											
Ph .	69	35	20	− 35	− 33	19	17	25	01	− 18	

Social, B—boss, E—exhibitor, M—unobtrusive mingler
Intellectual, E—empirical, A—abstract
Technical, C—clerical, M—mechanical, T—technical professional.
Creative, A—artistic, P—practical
Skill, Ph—physical, competitive; N—non-competitive has too few cases

tive interests with the two Intellectual classes are higher than the average inter-correlations either of the Social or the Technical classes of interest.

If our final analysis is correct, the grouping of "Social" interests, and of "Technical" interests into single categories covers up many actual differences. We cannot agree that the terms "Social," "Abstract" and "Mechanical" stand for three homogeneous, mutually independent, and collectively exhaustive categories of ability.

A VOCABULARY STUDY OF BIOLOGY NOTEBOOKS OF FIFTY REPRESENTATIVE SECONDARY SCHOOLS IN NEW YORK STATE

DON O. BAIRD

Associate Professor of Biology, Sam Houston State Teachers College, Huntsville, Texas

There have been many vocabulary studies made in recent years, especially in the field of elementary education Thorndike's "Teacher's Word Book" was a result of an extensive and comprehensive study of the vocabularies of elementary pupils and average citizens Recently Selke made a study of words in ten well-known textbooks in spelling in which he shows very clearly that the problem of vocabularies is by no means solved. Powers made a "List of Scientific Terms for High School Students," in which he lists the most important words from the field of natural science for pupils of the high school level Many others, as Bonser, Doran, Neher, Brandenburg, and Symonds, have contributed to research work in the various phases of vocabulary investigations

The study here presented was made to determine whether there is close agreement in the vocabularies used in biology textbooks and biology notebooks as written by high school pupils, and to compare the scientific terms used in such notebooks with those found in Powers "List of Scientific Terms for High School Students" and with those found in Thorndike's "Teacher's Word Book"

The writer collected biology notebooks from pupils in fifty representative secondary schools in New York State These notebooks were selected from the several classes in secondary school biology as good but not the best of each biology class The pupil vocabulary in these notebooks was examined to determine the following, extent of pupil vocabulary, extent of the use of scientific terms, the total number of words used, and the correct use and spelling of words

The total of all the separate words used in the notebooks was checked against the Thorndike list and the Powers list. In making up the list of words used in the notebooks proper names and symbols were counted in the total count for all the notebooks. The first step in discovering the extent of the pupil vocabulary as represented in the notebooks was a study of one of the smaller and one of the larger notebooks for comparison and to provide an initial word list for further study. Table I shows the results from the comparative study of the two notebooks.

TABLE I.—SUMMARY OF VOCABULARY STUDY OF ONE OF THE SMALLER AND ONE OF THE LARGER NOTEBOOKS

	Smaller notebook	Larger notebook
Number of words used but once	167	653
Number of words used one to four times	313	1,223
Number of words used five or more times	109	558
Total separate words used	422	1,781
Number of proper names	15	121
Number of symbols used	10	74
Number of abbreviations used	4	4
Total number of words used	2,220	15,249
Total number of words, including symbols, proper names, and abbreviations	2,250	15,448
Number of words not found in Thorndike's list	61	287
Number of scientific terms not found in Powers' list	7	68

The total number of scientific terms used in all the notebooks is six hundred forty-five. Many of the scientific terms were used but once in a single notebook. Comparatively few words were found to be misspelled and but few words "coined" by the pupils. Table II gives a summary of the vocabulary study of fifty-two of the notebooks.

TABLE II.—SUMMARY OF THE VOCABULARY STUDY

Total number of words used	488,820
Total number of separate words used	2,358
Total number of words not found in Thorndike's list	344
Total number of words not found in Powers' list	87

TABLE III.—WORDS IN THE NOTEBOOKS NOT FOUND IN THORNDIKE'S TEACHER'S WORD BOOK, OR IN POWERS' LIST OF SCIENTIFIC TERMS FOR HIGH SCHOOL STUDENTS

Absorption
Adaptation
Adenoids
*Advocacy
Aeration
Agar agar
*Ahead
Alimentary
Amoeba
Amphibia
*Anal
Anopheles
Antennae
Anther
Antibody
Aorta
*Apex
Aphid
Appendage
*Arterial
Artery
*Ashy
Assimilate
*Assimilation
*Astigmatism
Auditory
Auricle

Bacillus
Bacteria
Bacteriology
*Bag-like

TABLE III.—*Continued*

*Bailers
Bast
*Batting
Beaker
Beneficial
Bichloride
*Biconcave
*Bicuspid
Bile
Biology
*Blotter
Boll
*Box elder
*Bronchial
Burdock

Calyx
Cambium
Capillary
Capsule
Carapace
Carbohydrate
Carpel
*Carrion
Caudal
Cavity
Cellulose
Cephalothorax
Cerebellum
Cerebrum
Char
*Chitin
*Chlorate
Chlorophyll
Chloroplast
*Chyle
*Chyme
Cilia
Ciliate
*Cinnamon
Cloaca
Clot
*Coccus
Cochlea
*Cochineal
Cocklebur
Codlin
Combustion
Complex
Concave
Conservation
Convex
*Convolution
Coordinate
*Coordination
Cornea
Corolla
Corpuscle
Cortex
Cotyledon
*Coxa
Crayfish
*Cross bill
*Cross section
Cross-pollinate
Crucible
Crustacea
Cytoplasm

*Daddy-longlegs
*Decrease
Dehiscent
*Derangement
Devastate
*Dexterity
Diastase
Dicotyledon
Dioxide
Disinfectant
*Dissect
*Dissension
Dormant
Dorsal
Drupe

*Ectoplasm
Edible
Embryo
*Emulsion
Encyst
*Endoplasm
Endosperm
Enzyme
*Epicotyl
Epidermis
Erosion
Esophagus
Ether
Eustachian
*Every time
Excretion
Exoskeleton
Explode

Facet
Fahrenheit
*Fan-like
Fehling
Fermentation
Filament
Filter
*Flower-like
Focus
Follicle
*Fork-like
Formaldehyde
Fumigate

Gastric
Generalization
Generate
Germinate
*Glottis
Gram
Grippe
Gullet

Heliotropism
*Hessian
*Hibernation
Hydrochloric
Hydrogen
Hydrophobia
Hydroxide
Hygiene
Hygienic
Hypocotyl
Hypo-pharynx

Ichenumon
Ignite
Indigestible
Inorganic
Insecta
Insoluble
Integument

Katydid

Labium
Lacteal
Larva
Larynx
Lateral
Legume
Lengthwise
Lenticel
Liberate
*Licorice
Limestone
*Lysol

Maggot
Malaria
Mammal
Mandible
*Mandrake
Masticate
Maxilla
Medulla
Medullary
Membrane
Membranous
Merge
Mesothorax
Metamorphosis
*Metathorax
Metazoan
Micropyle
*Microsca
Microscope
Microscopic
*Mold
Mollusk
Monocotyledon
Monocotyledonous

Narcotic

TABLE III – *Continued*

Necessitate
*Neuron
Nicotine
Nitric (acid)
Nitrogen
Nitrogenous
Nodule
*Non-fatty
Nucleus
Nutrient

*One-celled
Operculum
Opium
Osmosis
Ovary
*Over done
Ovipositor
Ovule
Oxidation
Oxidize
Oxygen

Palisade
Palmate
Palpus
Pancreas
Pancreatic
Papilla
Paramecium
Parsnip
Pectoral
Pelvic
Pelvis
*Pericardium
Petiole
Pharynx
Photosynthesis
Physiologic
Pigment
Pinnate
*Pisces
Pistil
Placenta
Plasma
*Pliable
Plumule
Pollination
Pome
Posterior
Potassium
Protein
*Prothorax
Protoplasm
Protozoa
Pseudopodia
Ptyalin
Pulmonary

Quarantine

Rabies
*Rakers (Gill)
*Raphe
Reflex
Reproduce
Reproduction
*Respiration
Respiratory
Retina
Rhubarb
Rodent
*Rod-shape
Rootlet
Saliva
Salivary
*Salts
*Sawdust
Scavenger
Scientist
*Sclerotic
Sebaceous
Secrete
Secretion
Segment
Segmental
Self-control
*Semicircular
*Semifluid
Sensory
*Sheeting
Soluble
Spiracle
Spirillum
Spore
*Sputter
Sputum
Stamen
*Starling
*Steapsin
*Stegomyia
Sterilization
Stigma
Stimulant
Stimulate
Stimulus
*Stipule
Stoma
Stomata
*Stopper
Suffocate
*Sunlight
Swimmeret

Tachina fly
Tactile
*Taproot
*Telson
Testa
Thoracic
Thorax
Tibia
Trachea
Translucent
*Transparent
Transpiration
*Trichocyst
*Tricuspid
*Trillium
*Trochanter
Trypsin
*Tsetse
*Tweezers
*Typhoid

*Unglaze
Urea
Ureter

Vaccination
Vacuole
Ventral
Vertebrate
Vireo

Weevil

Xylem

* Not in Powers' list

Concluding, the writer would say it is evident that biology has a technical vocabulary more or less peculiar to it, and that the acquirement of this vocabulary by pupils is no easy task. There is a considerable range of scientific terms used in the several textbooks of secondary school biology and, as found in the notebooks, some of these terms are used but once by pupils. This would seem to indicate that the technical vocabulary of the textbooks may well be simplified, especially since many of the scientific terms are not found in Powers' "List of Scientific Terms for High School Students." The fact that relatively few words were misspelled or incorrectly used is noteworthy. It is also evident that a general vocabulary of less than two thousand words is sufficient for the average high school pupil in the study of elementary biology

BIBLIOGRAPHY

Bonser, F G, Burch, L H and Turner, M R. Vocabulary Tests as Measures of School Efficiency *School and Society*, Vol. II, pp 713-718

Baird, Don O · "A Study of Biology Notebook Work in New York State" Bureau of Publications, Teachers College, Columbia University, New York, N. Y, 1930.

Brandenburg, G C., Psychological Aspects of Language *Journal of Educational Psychology*, Vol IX, pp. 313-332.

Doran, E. W. A Study of Vocabularies. *Pedagogical Seminary*, Vol. XIV, pp. 401-438

Neher, H L: Measuring the Vocabulary of High School Pupils *School and Society*, Vol. VIII, pp 355-358.

Powers, S. R, Vocabulary of Scientific Terms for High School Students *Teachers College Record*, November, 1926

Symonds, P M: Size of Recognition and Recall Vocabularies *School and Society*, Vol. XXIV, pp. 559-560

Selke, Erich: A Study of the Vocabulary of Ten Spellers *Elementary School Journal*, Vol XXIX, pp 707-770

Thorndike, E L. "The Teacher's Word Book" Bureau of Publications, Teachers College, Columbia University, New York, 1921.

MECHANICAL "APTITUDE" OR MECHANICAL "ABILITY"?—A STUDY IN METHOD

O. L. HARVEY

In 1928 John Cox, a pupil of Spearman, stated the major findings on his doctorate dissertation* on mechanical aptitude as follows:

> The application of the tetrad-difference criterion to our specific correlation coefficients has failed to disclose the presence, here, of group factors . . . . The resulting specific correlation is adequately explained by a single factor running through those measurements in which this correlation occurs (130).

In a brief discussion of its relation to other "abilities," he pointed out that "mechanical aptitude" is to be strictly differentiated from "motor ability," which form of activity was "specially avoided" in his tests.

> Seeing, however, that the perception of space is largely developed by the aid of motor sensations, and that the material of the tests was spacial, it might appear at first sight that motor activity was an indirect determinant of success in dealing with this material. But the precise estimation of shapes, sizes, and distances, in which motor experiences might conceivably play a part, was not required in the tests. The essential processes on which success or failure turned were not perceptual but of the eductive kind in which motor experience could hardly have assisted (167–168).

In the initial definition of his problem, concerning the use of the term "aptitude," Cox wrote:

> Since the term "ability" must obviously extend its meaning to cover any kind of performance whatever, it would seem desirable to continue its employment in this wide sense, and to use the term "aptitude" when only the innate character of the "ability" is intended. Employing the terms in this way, we should say that a person actually able to carry out "mechanical" work has "mechanical ability," while one who has the appropriate innate mental constitution for acquiring this ability (whether he has actually done so or not) may be said to possess "mechanical aptitude" (40–41).

In these three key-statements, then, Cox makes his position clear, namely:

1. That he sought to discover, if possible, that innate "aptitude" which underlies the ability to carry out mechanical work;
2. That, as the result of experimentation, such an aptitude was found actually to exist;
3. That it is unique in its relation to "g" (Spearman's "general factor");
4. That it is a single group factor; and
5. That it is not the same as motor ability.

---

* Cox, J. W.: "Mechanical Aptitude." London, Methuen, 1928, pp. 210.

The instruments (tests) used by Cox consisted of mechanical models and diagrams of various sorts, which his subjects were required to explain or complete in one way or another, verbally. That is, the function of the subjects was essentially eductive, non-manipulative, non-motor Although it may be argued that, for eductive processes to be possible, certain implicit motor responses must have been involved, at least in the past acquisition of habits, nevertheless the actual test-situation involved no overt neuro-muscular behavior

Two years after Cox's work had been published there appeared a comprehensive study of mechanical ability* made possible by grants from the National Research Council Presumably Cox's work was unknown to the authors of this volume, at any rate until after their own study had been completed, for no reference is made to it. Their findings are, nevertheless, at least in their major reference, the same as his "As regards the uniqueness of mechanical ability," they wrote, "the evidence is consistently positive" (253). On the other hand (in contradiction to Cox's findings),

> mechanical ability . probably does not involve any single general factor. Low intercorrelations between different measures of mechanical ability suggest that factors of high specificity play a major rôle (300)

Mechanical ability they found to be "unique with respect to intelligence and also with respect to motor agility" (301), this last term being used synonymously with "motor ability," and being measured by means of parts of the Garfiel motor (gymnasium) test

The term "mechanical ability" was defined in this study to mean

> the ability to succeed in the actual manipulation of tools and materials and the ability to secure information about tools, and materials, and their uses (7)

It is to be differentiated from mechanical "capacity," (a term which, to the present writer, appears to be somewhat synonymous in this context with Cox's "mechanical aptitude"†)

The position of the co-authors of this group-study is thus made clear, as follows:

---

* Paterson D G *et al.*: "Minnesota Mechanical Ability Tests" Minnesota University Press, 1930, pages XXII + 586

† "Capacity," wrote the authors, "is always an inference from measured ability . For it is by no means evident that the development of mechanical ability may not be determined to a considerable extent by the presence and availability of tools and mechanical equipment" (7–8)

1. That they sought to discover whether there were a "mechanical ability," that is, the ability to succeed in the actual manipulation of tools and materials, regardless as to whether or not that "ability" be entirely attributable to innate capacity;

2. That, as the outcome of experimentation, they have revealed such an ability;

3. That it is unique in relation to verbal intelligence,

4. That it is, however, probably not a single group factor, but a constellation of factors of high specificity,

5. That it is unique in relation to motor agility.

The instruments (tests) selected as their final battery for the measurement of this "mechanical ability" comprize: (*a*) a test of steadiness, (*b*) a card-sorting test, (*c*) a block-packing test, (*d*) a form-board (spatial-relations) test, (*e*) an assembly test, (*f*) a test of mechanical knowledge (Stenquist's "*mechanical aptitude*"), and (*g*) a paper formboard test. Of these, the first three would appear, at least in the judgment of the writer, to be closely similar to tests constituting part of the Miles-Seashore "motor skills" unit;* whereas the spatial relations and assembly tests involve considerable manipulation.†

In so far as these two studies both claim to have isolated and discovered "*m*," they agree, in so far as the one clearly demonstrates "*m*" to be a single group factor, whereas the other equally clearly shows that very probably it is not a single group factor, but rather consists of a number of highly specific factors, they are in frank contradiction.

Professing to be no more than an interested onlooker in respect to the present discussions in this field, the writer has, nevertheless, ventured to seek explanations for this contradiction. He submits the following arguments, therefore, with considerable diffidence, hoping merely to attempt to clear up the differences which he finds.

DEFINITIONS, AND DELIMITATION OF FIELD OF INQUIRY

Possibly a significant reason for the difference in findings lies in the definition of the field of investigation. Although both studies claim to have revealed "*m*," actually Cox limits his investigation to the aptitude which makes mechanical ability possible, it excludes manipulative ability; it deals purely with the cognitive processes involved in the eduction of spatial relations The Minnesota study, on the contrary, clearly seeks to measure manipulative ability pure, the ability to use tools and materials. If, then, the ability to educe spatial relations is a function actually distinct from the ability to manipulate objects, it would appear evident that confusion has arisen out of the use of the designation "*m*" In which case, perhaps all that is necessary is to decide on some other symbolic designation for the factor of "mechanical ability" The real difference, however, appears to be more fundamental than one of definition alone, as will appear in the later discussion

---

* A study not included in the Minnesota bibliography, but reported in detail in *Journal of General Psychology*, Vol III, 1930, pp 38-65

† In this connection see criticism by Cox (pp 16-21, 166-168)

BASIS OF SELECTION OF TESTS

In line with the difference in definition of fields of investigation is the difference in basis of selection of tests. Cox decides against all available tests and constructs his own—tests involving, be it noted, no overt manipulative or motor activity on part of his subjects. Thus it could be argued that Cox's claim that "*m*" is not related to motor activity is substantiated by definition The Minnesota group is equally consistent, but strictly in terms of technique, not of philosophical orientation. Tests, roughly designated "mechanical," are correlated first with verbal intelligence and then with a shop-criterion, which, although of high reliability, is nevertheless, from the point of view of functional integrity, a somewhat ambiguous conglomerate Those tests which correlate as low as possible with intelligence and as high as possible with the criterion are selected to constitute the final battery—a basis of selection comparable, one suspects, with judgment in terms of the letter rather than of the spirit of the law

SPURIOUS AVERAGE INTERCORRELATIONS

Indeed it would appear that the nature of the tests determines the outcome. One actually expects Cox to discover marked consistency among his tests.* On the other hand, although the Minnesota tests are to be clearly differentiated from verbal intelligence, the fact that, taken as a group, they result in a low average intercorrelation does not necessarily mean that here is evidence of the absence of a large single group factor. An examination of Table XXXVI (page 233), "Intercorrelations of mechanical ability tests used in experiment proper, corrected for attenuation," yields an interesting observation suggestively in confirmation of this comment. Tests 1, 2, and 5 are as nearly as one could wish tests similar to those described by Seashore (*loc. cit.*) as purely motor tests; whereas Tests 4 and 7 are essentially "manipulative" and probably in part "eductive"—that is, in the sense in which these terms are used by Cox in the critical review at the commencement of his thesis. The average intercorrelation of all seven tests is .27. But, when taken alone, the average intercorrelation between the three motor tests is as low as .23* while the correlation between the two tests involving manipulation is actually as high as .63. And a similar argument applies in respect of all of the tables submitted on pages 432–433, 235, 238, and 232.† Intercorrelate the tests which are, as nearly as one can judge, purely motor; intercorrelate, again, those which are both manipulative and eductive, exclude altogether those which fall into neither category, and consistently *one discovers a high average intercorrelation for the manipulative* tests, while the motor tests intercorrelate even lower than the average for the entire table as given by the authors, and in each case the range of correlations is narrower. Does not this suggest that possibly the reason for low intercorrelations, the high specificity, discovered by the Minnesota group is the fact that most of the tests used by them are actually little more than motor tests, and that their "*m*" is really pretty much the same "factor" as that which Seashore claims to measure—a constellation of highly specific motor skills? May we not look here for the fundamental difference between Minnesota's "*m*" and Cox's?

---

* It is not necessarily implied by this argument that Cox's "*m*" factor includes all available tests of "*m*" The denotation of a group factor, as the writer understands it, is a matter of proof, which Cox has not submitted Nevertheless, so far as he has gone, his tests consistently agree

A priori there would appear to be four phenomena which one should distinguish in dealing with mechanical activities:

1. Specific motor skills per se (as indicated by Seashore);
2. Manipulative or mechanical ability (which the Minnesota group sought to measure. It is not, according to Cox, the same as the "practical ability" suggested by Macfarlane),
3. Mechanical aptitude (an intellective activity involving the ability to educe space relations between objects, without necessitating their prehension and manipulation); *and finally,*

---

* *Cf.* Seashore's uncorrected intercorrelation of .25.

† It is appreciated that in every case there are only two manipulation tests to correlate. The suggestion contained in this argument is nevertheless sustained.

4 The sheer intellective factor, labelled "*g*" by Spearman, which presumably pervades all cognitive activity, including the mechanical

The Minnesota group seems to have got some of these concepts confused From the point of view of internal consistency its battery would appear to be far superior to Cox's. In so far as its best battery of tests (not only mechanical) correlates high (about .8) with a criterion of excellent reliability (about .95) it is probably perfectly satisfactory as a measure of skill in shop-work performed by junior high school boys. The study is splendidly watertight But, in view of the above arguments, one wonders whether its assumptions are entirely warranted. Cox's position, at least in theory, appears to be the more securely tenable His tests are consistent in that they deliberately exclude motor and manipulative activity in their immediate solution; and in that, before finally determining the organization of "*m*," Cox eliminates (or claims to have eliminated) the influence of "*g*" from his tests.

On the other hand, Cox's material is vague and often exceedingly unsatisfactory. His populations are heterogeneous (by comparison with those of the group study) and most of them too small The same tests have not been used on all groups alike. And one is left with an uncomfortable suspicion that success in answering his tests (no scoring procedure has been indicated) depends to such an extent upon verbal intelligence, that "*m*" (even when, to all intents and purposes, the influence of "*g*" has been eliminated) still contains a powerful amount of "*g*."

It might be an interesting piece of research to analyse in terms of the tetrad difference technique the data obtainable from a homogeneous group in response to (*a*) a test of verbal intelligence, and (*b*) a combined battery including tests such as (1) the Minnesota assembly and spatial relations tests (for "mechanical ability"); (2) a selected group of Cox's models, explanations, etc. (for "mechanical aptitude"), and (3) either the Minnesota tests for steadiness, card-sorting, packing blocks, and tapping, and Link's machine operators' test, or some form of motor skills battery, similar to that developed by Seashore (for "motility").

Cox and the Minnesota group have demonstrated how such an investigation can be conducted Possibility an experiment along the lines here indicated might help to clear up the basic issues involved in their major disagreement.

THE RELATIVE EFFORT OF CHILDREN OF NATIVE *VS.* FOREIGN BORN PARENTS*

S. EDSON HAVEN

Ohio State University

INTRODUCTION

The problem here attempted is to determine by objective means the relative effort of children of native *vs.* foreign born parents. The interest in this study grew out of recognition of the tendency to discriminate against the foreigner in the manner which he applies himself in accomplishing school tasks. The significance of such an experiment may be more fully appreciated when we stop to consider that approximately one in every eight, or thirteen per cent, of the population of the United States is of foreign birth.

The AQ, or Accomplishment Quotient, is used as the index to effort The AQ is the EQ (educational quotient) divided by the IQ (intelligence quotient). Quoting from the Twenty-First Yearbook,[1]† "The IQ is a measure of the native ability of the child and shows his potential rate of progress. The EQ is a measure of the educational attainment of the child and shows his actual rate of progress. The Accomplishment Quotient is the degree to which his actual progress has attained to his potential progress by the best possible measure of both. It is a mark which evaluates the accomplishment of the child in terms of his own ability." Franzen[2] states, "Since the IQ is the potential rate of progress and the EQ is the actual rate of progress, the ratio of the EQ to IQ gives the percentage of what the child could do, that he has actually done This would be a mark of the child's effort, a mark of concentration and interest that the child has in school work, and so far as no other inherited traits than intelligence affect school work it is a measure of the efficiency of a child's education thus far. If there are such other innate bases it is also a measure of inherited traits, such as concentration and effort At any rate, it is a measure of the child's accomplishment, and so of the effort and concentration as they really are at present working under those school conditions" Witty and Lehman[3] defining the AQ say, "The AQ is a comparison of the EA (educational age) to the MA or the EQ to the IQ Deviations above and below 1.00 indicate the extent to which effort is expended and ability exercised in the tasks of the school. Presumably an AQ of a 1.00 indicates that the mental ability and the educational attainment of the child or group are developing coordinately."

---

* The writer is indebted to Professor James P Porter of Ohio University for guidance in this study

† Numbers indicate reference in bibliography

Goodenough[4] reminds us that immigrant groups and their immediate descendants differ markedly in their performances on the ordinary type of intelligence test She suggests two theories to account for these differences The first ascribes the inferior showing to such postnatal factors as inferior environment, poor physical condition, and linguistic handicaps. The second recognizes to some degree the effect of these factors on test-results, but holds that it is impossible to account for all the facts which have been observed upon any other hypothesis than that of innate differences. She goes on to say, "Although a test is completely independent of language, the rank order of various racial groups corresponds very closely to the results of other investigations using verbal tests"

Many experiments have been undertaken and reported relative to the linguistic handicap with which the foreigner is faced when he takes a test designed to measure all children but has been standardized on English speaking groups.

Brown[5] has made an investigation on retardation in several schools in northern Michigan. He found there was a wide variation in retardation among the children of different nationalities. It was revealed that after a pupil of foreign born parents had attended an American school for one or two years, he tested as high by employing the English language as by using the native tongue

Young[6] in a study carried on in the San Jose schools in which he used the Army Alpha and Beta Tests on American, Italian, Portuguese, and Spanish Mexican children, found that the Alpha Test proved a better measure of the Latin groups than the Beta He concluded that language handicap was not the cause of school difficulty.

Jones[7] made a comparison of the scores on the Myers Mental Measure and the McCall Multi-Mental Scale, discovering there was no appreciable difference between the scores made by the natives and those made by the foreign. It is contended that the former test is largely non-verbal while the latter is highly verbal

A statement from the Twenty-Seventh Yearbook by Burks and Kelley reads,[8] "We find studies that purport to measure the effect of language handicap on verbal intelligence test scores by comparing the mental ages of foreign children earned on verbal and non-verbal intelligence tests. The mental ages of children of certain low testing nationalities commonly turn out to be closer to the norms of American children when measured on non-verbal tests than when measured on verbal ones But in as much as verbal and non-verbal test scores, even for American children, seldom correlate with one another higher than .6 or .7, it is obvious that, although both types of test are called 'intelligence' tests, they measure about as much not held in common as they measure of what is held in common. Hence, it is not legitimate to infer from such data alone that language handicap accounts for the low scores of the foreign children on verbal tests."

Although further investigations will be necessary to settle the question, the foregoing citations have some significance May it not be reasonably assumed that some personality trait such as effort is playing an important rôle?

METHODS AND PROCEDURE

The schools selected for making this study are located in mining communities of Athens County, Ohio Approximately ninety-five per cent of the workers are employed in the mines. Under normal conditions the economic situation is generally good. Due to a strike which has lasted for over a year a severe depression has developed. *For the most part, foreign standards of living prevail* The church and school are the chief sources of culture.

The groups under consideration consist of the children from the fourth to the eighth grade See Table I for a detailed statement of nationalities Those whose parents are foreign born constitute the one group and those children whose parents are native born make up the other group All children were born in America, but the majority hear a foreign language at home The pupils of both groups are of much the same social status Arlitt[9] emphasizes the importance of taking this factor into account She finds there is a marked difference in the distribution of intelligence in groups of children of the same *nationality but of a different social level*

It was recognized at the outset that there was a selected group of natives. Goodenough[1] remarks that a person of low intelligence tends to gravitate to those neighborhoods where the economic requirements are minimal; and, once there, he reacts toward his surroundings along the line of least resistance From personal observation and inquiry it was learned that no prejudice or favoritism existed on the part of the teachers. The native and foreign were treated alike.

The hearty cooperation of the teachers and principals was had throughout. The pupils manifested a great deal of interest in the tests by giving close attention to all instructions and suggestions of the examiner.*

The tests used were the Otis Classification Forms A and B, and special tests of the sentence completion and auditory memory type

TABLE I.—DISTRIBUTION OF NATIONALITIES

NATIONALITY	NO. OF CASES
School A	
American	65
Hungarian	23
Polish	24
German	6
French	13
Italian	3
Slavish	4
Lithuanian	2
Foreign	75
Native	65
Total.	140
School B	
American	98
Hungarian	45
German	4
Polish	1
Foreign	50
Native	98
Total	148
School C	
American	68
School D	
American	43

---

* Most of these pupils were accustomed to taking standardized tests. Form A of the National Intelligence Test and Forms A and B of the Stanford Achievement Test were given the preceding year in an effort to determine the influence of special intensive training in various school subjects. While the results, when considered from the point of view of this study tend to confirm the findings here reported it has seemed best to omit further mention.

The general procedure was to administer the first form of the test, then after a period of a few weeks give the second form of the same test

The problem naturally divides itself into two distinct approaches, each of which will be considered separately.

TEST FINDINGS—PART I

In attempting to determine effort by using the AQ as an index, it was decided that the Otis Classification was the best examination available, in that, it is composed of both the so-called educational and mental ability tests. The first form was given in January and the second in April. The AQ for each individual on each test was derived in the way previously stated Only results from those pupils receiving both forms are considered The different grades in Schools *A* and *B* were highly comparable in numbers of each sex, and in relative numbers of foreign and native

TABLE II.—THE MEDIAN EQ, IQ, AND AQ FOR THE FOUR SCHOOLS ON FORM A OF THE OTIS CLASSIFICATION TEST, ALSO QUARTILE RANGE FOR THE AQ

School	EQ	IQ	AQ	Quartile range of AQ
C	82.77	86.81	95.35	7.08
D	82.91	80.83	96.94	6.02
A (Native)	86.78	86.54	99.42	6.88
A (Foreign)	84.5	74.8	103.28	7.39
B (Native)	79.54	81.59	94.02	6.49
B (Foreign)	83.12	81.30	100.83	7.06

The reason for giving the two forms was for the purpose of checking the consistency of the two groups in achieving school work within the same limits of time

Schools *C* and *D* were selected and tested to see if the absence of the foreign affected the academic achievement of the native as determined by these tests.

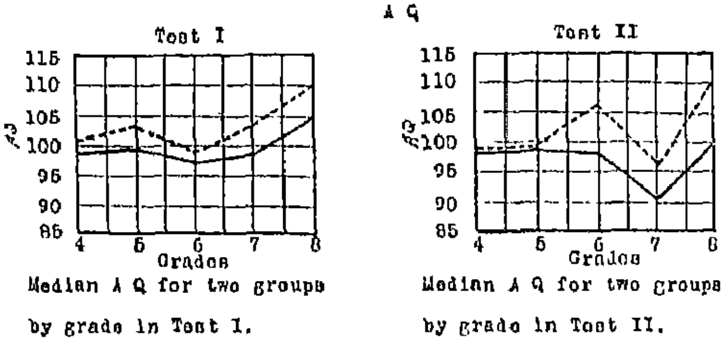
Table II gives the median EQ, IQ, and AQ, for the different schools and groups The quartile range is also given for the AQ In Table III are the same data on the second test.

In Fig. 1 and 2 are given the curves for the median AQ on the two tests for the two groups by grade

Table III.—The Median EQ, IQ, and AQ for the Three Schools on Form B of the Otis Classification Test, also Quartile Range for the AQ*

School	EQ	IQ	AQ	Quartile range of AQ
D	80 83	84 04	03 80	7 03
A (Native)	88 75	02 75	07 5	6 69
A (Foreign)	88 05	87 81	100 08	8.14
B (Native)	84 37	88 84	05	6 02
B (Foreign)	88 57	87	102 5	7 75

* As School *C* had dismissed for the summer the second test could not be given



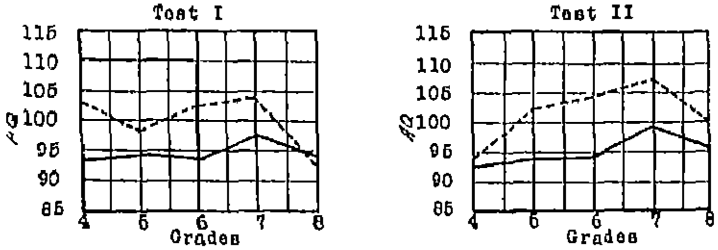
Median A Q for two groups by grade in Test I.

Median A Q for two groups by grade in Test II.

------ Foreign

——— Native

Fig 1.—School A.



Median A Q for two groups by grade in Test I.

Median A Q for two groups by grade in Test II.

Fig 2.—School B.

Tables IV and V give the percentage of foreign reaching or exceeding the median of the native. In other words, the figures represent how many foreign are as good as or better than half of the native "

If the percentage is greater than sixty or less than forty it may be considered significant. Starch[12] states that any differences lying between forty per cent and sixty per cent of the number of either group reaching or exceeding the median of the other are practically negligible. If sixty per cent of one group reach or exceed the median of the other it means that ten persons in a hundred of the one group, are by a small amount superior to the other.

TABLE IV.—THE PERCENTAGE OF FOREIGN REACHING OR EXCEEDING THE MEDIAN OF THE NATIVE ON FORM A OF THE OTIS CLASSIFICATION TEST

School *A*

Grade	4	5	6	7	8
EQ	33.3	65	23	47	50
IQ	33.3	65	30.7	35.2	10
AQ	73.3	60	76.9	82.3	70

School *B*

EQ	75	66.6	46	75	33.3
IQ	100	80	23	50	33.3
AQ	87.5	86.6	84.6	75	83.3

TABLE V.—THE PERCENTAGE OF FOREIGN REACHING OR EXCEEDING THE MEDIAN OF THE NATIVE ON FORM B OF THE OTIS CLASSIFICATION TEST

School *A*

Grade	4	5	6	7	8
EQ	40	65	30.7	41.1	60
IQ	26.6	75	38.4	17.6	10
AQ	66.6	50	60.2	82.3	80

School *B*

EQ	87.5	80	61.5	50	33.3
IQ	75	66.6	38.4	50	16.6
AQ	87.5	100	84.6	62.5	66.6

TABLE VI.—THE PERCENTAGE OF FOREIGN REACHING OR EXCEEDING THE MEDIAN OF THE NATIVE ON FORMS A AND B OF THE OTIS CLASSIFICATION TEST

	Form A	Form B
All foreign in school *A*.		
EQ	49 3	57 3
IQ	14	45 3
AQ	80	60 0
All foreign in school *B*.		
EQ	68	74
IQ	56	58
AQ	86	68
Foreign in *A* and *B*:		
EQ	41 6	58 4
IQ	52.	42 4
AQ	70	07 2

PART II

The central problem is here approached from a different angle than the preceding experiment Special tests were devised after Sentence Completion and Auditory Memory Tests of Otis. The first type of test is made up of sentences from a given story with words or phrases omitted. It is the pupil's task to fill in these blank spaces with appropriate words that he may have a rather complete story when he has finished. A list of possible words follows each set of seven or eight sentences For example: Once   was a terrible giant ogre.

pleasantly
their
story
died
fiercely
crawl
terribly
lived
there
sitting
built
pass

The best word is *there*, consequently it is placed in the blank space It will be recalled that the second type of test depends on the ability of the pupil to retain and recall events mentioned in a story which he hears read. For example: Was Gotham known for its wondrous wise men? (yes, no, didn't say). The correct response is "yes" according to the story, therefore this word is underscored.

There were two sets of these examinations: the first set was given on April 12 and the second on April 26.

TABLE VII.—MEDIAN SCORES FOR THE NATIVE AND FOREIGN ON THE SENTENCE COMPLETION AND AUDITORY MEMORY TESTS

	Foreign		Native	
No. of examination	1	2	1	2
Completion tests	12.22	14	15.10	14.25
Memory tests	14.00	15.38	15.53	14.72

Percentage of foreign reaching or exceeding the median of the native

No. of examination	1	2
Completion tests	35.2	41.1
Memory tests	73.5	73.5

The Completion and Memory Tests had thirty-five questions or statements each. The score was the number answered correctly.

At the close of the first test the pupils were told that in two weeks the examiner would be back to give them a similar set of questions and incomplete sentences. They were advised that the stories from which these tests would be compiled were to be found in specified books in their library, and that from the six stories the names of which

TABLE VIII.—MEDIAN SCORE FOR THE NATIVE AND FOREIGN ON THE TWO EXAMINATIONS IN THE SEVENTH GRADE

	Foreign		Native	
No. of examination	1	2	1	2
Completion tests	12	14.16	17	17.5
Memory tests	13.85	16.25	16.25	16.25

Percentage of foreign reaching or exceeding the median of the native in the seventh grade

No. of examination	1	2
Completion tests	37.5	62.5
Memory tests	43.7	62.5

were placed on the board, two would serve as the basis of the second examination. No further suggestions were given. The teachers were requested to make no further mention of the work.

Table VII gives the results for the two groups in the two examinations Table VIII gives the results for the two groups on the two examinations for the seventh grade, the reason for selecting this particular grade being that there were the same number of the foreign and native

DISCUSSION AND CONCLUSIONS

When Colvin[13] compared the scores earned, on the Otis Group Intelligence Scale, Forms A and B, by children of Brookline, Massachusetts and Cincinnati, Ohio he found the latter to be inferior He attributed this inferiority to differences in opportunities to learn words and acquire skill in their use He supported his conclusion with the fact that the children in the poorer localities in Brookline did not score as high in the entire test as did the children in the more favored localities In the Arithmetic test "largely non-verbal in its nature" the scores were not inferior to those made by the children in the better sections.

The results of our own study would seem to support Colvin's contention. The scores on the entire test were lower than the established norms However, upon a careful examination of the Arithmetic problems of the Otis Classification it was found that both groups did equally well. In view of this fact it is apparent that both groups were competing on an approximately equal basis with reference to the language factor. In no case was it discovered that the foreign were more handicapped linguistically than the native

From Tables II and III it may be observed that in all cases the foreign make a higher median AQ than the native. It is true that the native make a higher IQ but the difference is so slight that we can attach little, if any, significance to it

From these results it is evident that the native are more highly selected than the foreign, the quartile range in AQ is greater for the foreign group. Inspection of frequency distribution tables also shows greater variability The selection of the native is emphasized by the fact that nearly fifteen per cent of those taking the first test had moved from the community by the time of the second This accounts for the limited number of cases in the study, as we considered only those pupils who were tested both times

The schools in which no foreign were found, tested on a par with the other two schools. That there were no foreign in Schools *C* and *D* was learned from questionnaires filled out by the pupils and confirmed by the teachers An attempt was made to get information on the different grades with reference to median AQ. Figures 1 and 2 show a great deal of fluctuation, but it will be noted that the foreign and native groups align themselves in curves exhibiting marked consistency.

The eighth grades of Schools *A* and *B* were very small Table IV and V present data which might lead one to conclude that the low IQ of the foreign was responsible for their high AQ. Where this difference occurs the probable explanation is the smallness of the class.

Table VI shows the percentage of foreign reaching or exceeding the median of the native to be at least significant.

In approaching the problem by the second method, the Special Completion and Memory Tests, it was assumed that the group showing the most improvement would be putting forth the most effort. This assumption could be reasonably justified on the basis of the preceding experiment. Tables VII and VIII would indicate the foreign to be superior. Interesting details in connection with this second experiment are: Of seven who were effortful enough to read the stories for the second examination, five were foreign; the correlation between the two examinations was 794 with a PE of 032

Our study may have something in common with an investigation which has been carried on by Miller[11] with freshman engineering students at the University of Michigan "The low students are, in the main, children and grandchildren of American-born men and women. They also are the sons of the better educated and more prosperous parents. A surprisingly large percentage of the high students are the grandsons of foreign-born men and women They are the sons of parents having little education or none In the main they are the sons of parents whose education does not include the high school Twenty-four per cent are the sons of parents who had less than a grammar school education" He further states. "Low students are lacking in the quality of vision, powers of analysis and coordination. They are superficially bright and ingenuous, but their tenaciousness of purpose and capacity for sustained mental effort are so low that their curve of effort has a steadily downward trend"

May not the above discussion and our findings suggest a possible and partial explanation for the prevailing opinion of the thrift and rapid economic rise of the foreigner?

Some conclusions to be drawn from this study may be briefly stated as follows:

1. Groups of children of foreign-born parents are consistent in making higher median Achievement Quotients than are groups of children of native-born parents

2. In most cases the median Intelligence Quotient is slightly higher for the groups of native children than for the foreign.

3. Both groups were handicapped in the use of language as measured in all the tests.

4. The foreign have little or no inhibiting or unfavorable influence on the school progress of the native.

5. By the use of objective tests the foreign prove themselves superior to the native in effort The relative significance of this personality trait is at least suggested by the manner in which the foreign overcome obstacles in making new adjustments

Bibliography

Arlitt, Ada H (9): On the Need for Caution in Establishing Race Norms. *Journal of Applied Psychology*, Vol V, No 2, pp. 179-183.

Blackwood, Beatrice. A Study of Mental Testing in Relation to Anthropology. *Mental Measurements Monograph*, Serial No. 4, Dec , 1927

Brigham, Carl C : "A Study of American Intelligence." Princeton University Press, 1923.

Brown, Gilbert (5). Intelligence as Related to Nationality *Journal of Educational Research*, Vol V, No 4, pp. 324–327.

Burks, Barbara and T L Kelley (8): Statistical Hazards *Twenty-Seventh Yearbook*, p 20

Colvin, S S and R D. Allen (13) · Mental Tests and Linguistic Ability *Journal of Educational Psychology*, Vol. XIV, pp. 1–20.

Dickson, Virgil E · "Mental Tests and the Classroom Teacher " World Book Co , 1924.

Franzen, Raymond H (2) The Accomplishment Quotient, A School Mark in Terms of Individual Capacity. *Teachers College Record*, Vol. XXI, pp. 432–440

Goodenough, Florence (4–10)· Racial Differences in the Intelligence of School Children. *Journal of Educational Psychology*, Vol IX, pp. 388–397

Hankins, F. H · "The Racial Basis of Civilization " A. A Knopf, 1926

Hirsch, Nathaniel: An Experimental Study of the East Kentucky Mountaineer *Genetic Psychology Monographs*, Vol. III, No. III

Jones, V. A (7): A Study of Non-Verbal Nature and Validity of Myers Mental Measure *Journal of Educational Research*, Vol. XVI, No III, pp. 203–209.

Mead, M. Group Intelligence Tests and Linguistic Disability among Italian Children *School and Society*, Vol. XXV, No. 642, pp 465–468.

Miller, H W (14): Profits Derived from Segregating College Students on the Basis of Ability. *Science*, Vol LXV, pp. 427–429

Pintner, R (1). *Twenty-First Yearbook* Pp 165

Pintner, R : "Intelligence Testing," Chapter 18 Henry Holt, 1923

Pintner, R and R. Keller. Intelligence Tests for Foreign Children *Journal of Educational Psychology*, Vol XIII, pp. 214–222.

Pintner, R. Comparison of American and Foreign Children on Intelligence Tests *Journal of Educational Psychology*, Vol XIV, No V, pp 292–295.

Poffenberger, A T (11) "Applied Psychology Its Principles and Methods" D. Appleton and Company, 1927

Sereta, K. E A Comparative Study of 100 Italian Children at the Six Year Level *The Psychological Clinic*, Vol XVI, No. VII, pp. 216–231

Starch, Daniel (12): "Educational Psychology." Macmillan Co, 1927, pp 76

Witty, P. A and H. C Lehman, (3): The Rise and Fall of New Educational Methods. *Journal of Educational Methods*, Vol VII, No. I, pp. 2–6.

Young, Kimball (6): Mental Differences in Certain Immigrant Groups. *University of Oregon Publications*, Vol I, No. XI.

THE EFFECT OF THE FORM OF A COMBINATION IN THE LEARNING OF A MULTIPLICATION TABLE BY BRIGHT AND DULL CHILDREN

F. T. WILSON

State Teachers College, Buffalo, New York

In an experimental study reported elsewhere,[1] selected groups of fifteen children each, learned a multiplication series from 2 × 67 to 9 × 67, inclusive. Practice was provided by a card, upon which the eight combinations appeared in both vertical forms, that is as $\begin{array}{r}67\\ \underline{2}\end{array}$ and $\begin{array}{r}2\\ \underline{67}\end{array}$, with the answers given at the bottom of the card in both horizontal forms, that is, as 2 × 67 = 134, and 67 × 2 = 134 The children read the combinations above orally, and found the answers by looking below and giving them orally. The subjects were selected by individual test with the Stanford Revision of the Binet and by chronological age. Half of them were practically nine years old and half twelve. Half of each of these ages were of IQ's between eighty and ninety, and the others between one hundred ten and one hundred twenty Full description of the subjects and the technique of the experiment are given in the reference cited above

The table shown herewith gives the record of the correct answers of each group A striking likeness is seen in nearly every pair of the combination forms. In only four cases are the differences in the scores of pairs more than five. Investigation of the detailed records indicates that these are purely chance results, explainable in the small number of responses in certain instances and the cumulative effect of very few subjects in small scores

The main point of interest in these data seems to be not so much that, in learning the answers, the two forms of an arithmetic combination are of no particular difference in difficulty, as that there seem to be no distinguishing differences in that regard between children of rather low mental ability and others of rather high mental ability In effect it seems to suggest that the learning process, whatever it may be in ultimate elements, is much the same for dull and bright. This is supported by other evidence in the table not stressed in this report because it has been treated more extendedly elsewhere, namely, that the combinations which were hardest for the bright were also hardest for the dull, and that the combinations easiest for the dull were also strikingly easiest for the bright.

---

[1] Wilson, F. T.: "The Learning of Bright and Dull Children" Teachers College Contributions to Education, No. 292, 1928

These points tend to turn the interest of those concerned with the guidance of learning processes of human beings to two concepts said to be at present of some psychological repute. They are, first, that human beings, despite great variations in abilities and interests, learn in much the same way one to another. There are differences in rate, accuracy and retentiveness of learning, but basically there seem to be conditions and progress of like kind The second concept is that a learning task such as this is attacked and progress made in it in a way which suggests the unity or wholeness of the entire matter, rather than a separateness of parts Grasping the plan of the task, as expressed in an answer, the mind seems not to be disturbed, perhaps, by the arrangement of parts, but seeks to handle the material as a whole

TABLE I —NUMBER OF CORRECT RESPONSES TO EACH FORM OF THE COMBINATIONS BY THE DULL AND BRIGHT GROUPS

Combinations	Dull 9	Bright 9	Dull 12	Bright 12	Total
2 × 67	21	93	99	118	331
67 × 2	22	98	92	118	330
3 × 67	31	74	73	102	280
67 × 3	20	74	71	101	266
4 × 67	1	14	2	20	37
67 × 4	0	17	3	17	37
5 × 67	43	85	76	101	305
67 × 5	44	82	72	110	308
6 × 67	28	63	35	56	182
67 × 6	28	62	43	61	194
7 × 67	0	13	10	21	44
67 × 7	0	13	9	23	45
8 × 67	7	44	23	58	132
67 × 8	7	49	18	53	127
9 × 67	29	95	58	89	271
67 × 9	31	99	61	86	277
Total	312	975	745	1134	3166

STUDY HABITS OF TEACHERS COLLEGE STUDENTS

HUGH M. BELL

State Teachers College, Chico, Calif.

The question of the amount of time given to study is of importance to both the student and to the administrator. Students in a teachers college, preparing to guide others in learning, need especially to demonstrate to themselves the relationship which exists between their study habits and their scholastic achievement. From the administrator's standpoint anything approaching successful student guidance cannot be attained without definite information in regard to study activity and the relation it bears to intelligence and scholastic standing.

In this investigation we sought to answer these questions: When do students study? How long do they study? What is the nature of their study activity? How is the amount of study time related to intelligence and scholarship?

Students in General and Educational Psychology classes were used as subjects. The investigation was conducted during the school year 1928-1929. A total of one hundred twenty-seven students kept daily records of their study time for twenty-eight days. Recording was not begun until after the third week of each semester. Keeping the records was considered a part of the regular work of the courses. Every effort was made to secure the cooperation of the students since this type of investigation depends so much upon it. The instructor took one-half hour of the period on the day that the record blanks were given out to explain the purpose of the investigation and to show them the importance of keeping the records carefully. Each succeeding week when new blanks were given out the students were again urged to keep the records faithfully.

The study record consisted of a specially prepared sheet on which columns were provided wherein the student recorded each day's study activity. The students were given sufficient record blanks for one week. At the end of the week they turned in their records and got a new set of blanks. Columns were provided in which the students could record the following information: Time of study —the day of the week and whether it was done in the morning, afternoon or evening, type of study—whether reading in a textbook, outside reading assignments or written work such as themes and problems.

The study day was divided into three periods: Morning, from time of arising until twelve noon, afternoon, from twelve noon until six o'clock; evening, from six o'clock until time of retiring.

During the school year all the students were given the American Council of Education Psychological Examination Reports were secured from the registrars office of the grade points earned by each student in all subjects for one semester. Correlations were computed between study time, scholarship and intelligence.

RESULTS

From analysis of the 3556 days of study reported by the group, as shown in Table I, there was a total of nearly 395,000 minutes, 6583 hours, given to study by the entire group over the period of four weeks. This would be an average of one hundred eleven minutes, or just short of two hours per day for each student. Twenty-four per cent of the total study work was done in the morning period, thirty-two per cent in the afternoon, and forty-four per cent in the evening. The results indicate that nearly one-half of the students' studying is done in the evening by artificial light The average number of minutes per student per day for the morning period is twenty-seven minutes, for the afternoon thirty-five minutes, and for the evening period forty-nine minutes These averages assume that all students studied on each of the twenty-eight days that the investigation covered

TABLE I—TIME OF DAY WHEN STUDY WAS DONE

Time of day	Number of minutes	Per cent	Average number minutes per student, per day
Morning	96,493	24	27
Afternoon	125,167	32	35
Evening	173,348	44	49
Total	394,908		111

A comparison of the study activity on the various days of the week is recorded in Table II. The number of minutes recorded for any given day is the sum of all the study time reported by all the students for that day. There is a gradual increase in the amount of study from Sunday to Tuesday where it begins to drop off to Thursday. From Thursday to Sunday there is a pronounced drop The most outstanding drop is from Thursday to Friday when less time was given to study than on any other day of the week The difference between Saturday and Friday is not as much as one might have supposed.

In the second column of this table the average number of minutes per day for each student is reported In these averages the actual number of students studying on a given day is taken into consideration In Table I where the number of students studying on a given day was not taken into account, it appeared that, on the average, our students were studying one hour and fifty-one minutes per day. In Table II those who did not study were eliminated from the averages and the mean increased to two hours and thirteen minutes.

TABLE II.—COMPARISON OF STUDY ACTIVITY ON VARIOUS DAYS OF THE WEEK

Days of week	Number of minutes	Daily average, number of minutes per student	Number of students studying	Per cent
Sunday.	43,890	128	343	67
Monday	66,733	140	476	94
Tuesday	83,488	168	505	99 4
Wednesday	72,881	148	488	96
Thursday	69,293	140	490	97 6
Friday	33,496	92	374	73 6
Saturday	25,217	116	213	41

The number of students studying on a given day is reported in column three of Table II. This represents the total for the four week period Thus if all the students had studied on a given day the Number of Students Studying would have been five hundred eight since there were one hundred twenty-seven students who kept records for four weeks. In the fourth column the per cent of students studying on any given day is shown In general the number of students studying shows about the same tendency to increase or decrease that the number of minutes shows However, there are interesting exceptions to this. The number of students engaging in study increased from Wednesday to Thursday, but the duration of the study periods decreased. There is a greater number of students studying on Friday than on Sunday, but the total number of minutes given to study on Sunday is twenty-four per cent greater than that on Friday There is a general tendency to decrease the length of the study period as the week-end approaches. A comparatively small number of students, forty-one per cent, are studying on Saturday, however, their study activity seems to be of a better quality than Friday's in as much as it is carried on for longer periods of time. Both from the standpoint of the number of students studying and the average length of time given to study, Tuesday tends to be the best day of the week for study.

Book,[1] in a recent investigation of reading among college students states that it has been estimated that the average student gets ninety per cent of his information in college from some type of reading activity. In Table III evidence is presented which apparently substantiates this estimate. Not all the students signified the type of study, whether reading or writing, and hence our data represent the reports of only those who did. The total number of minutes for those students who signified the type of study was 159,257. Of this, sixty-three per cent was given to reading textbooks; nineteen per cent to reading collateral material, 8.5 per cent to working problems; and 9.5 to writing papers and themes. Eighty-two per cent of all the study time was used in some type of reading activity and eighteen per cent in writing work. This is, of course, only a rough estimate since some types of study are difficult to classify under the above headings.

TABLE III.—COMPARISON OF AMOUNT OF TIME GIVEN TO READING AND WRITING IN STUDY ACTIVITY

Reading activities				Writing activities			
Textbook		Collateral		Problems		Papers	
Number minutes	Per cent	Number minutes	Per cent	Number minutes	Per cent	Number minutes	Per cent
100,886	63	30,530	19	13,552	8.5	14,280	9.5

Since some students carry heavier scholastic loads than others it was necessary to devise a means of equalizing this difference before determining the average study time for a given student to be used in our correlations. This was accomplished by finding the average number of minutes per week spent in study and then dividing this by the number of units or hours carried for the semester. This gave the number of minutes studied for a given hour of school work. Courses which did not require preparation, such as physical education activities, were eliminated. When the student's grades and study time were correlated the Pearson r was +.317 ± .05. The coefficient, when grades were correlated with intelligence, was +.565 ± .04. Intelligence correlated with study time +.003 ± .06. The first two coefficients are approximately the same as secured by May[1] in an earlier study of this same character. However, he secured a negative correlation between intelligence and study time. Using the partial correlation method and holding intelligence constant, the correlation between study time and scholastic standing is +.382 ± .05.

---

[1] Book, Wm. F., How Well College Students Can Read. *School and Society*, Vol. XXVI, August, 1927, pp. 242–248.

From these coefficients it would seem that the intelligence test was a better indicator of scholastic success than study time. Intelligence is not definitely related to study time since the student with a high or low score is as likely as not to study a long or short period of time. This is illustrated by reference to an individual case. Two students who stood at the ninety-eighth percentile in the intelligence test studied, on the average, seven minutes, and four hours and nineteen minutes per day respectively. The student whose study time was seven minutes per day made a C average in her courses. The other student had a straight A average.

SUMMARY

An investigation was carried on with one hundred twenty-seven teachers college students of freshman to senior standing to analyze the time factor in study and to relate it to intelligence and scholastic standing. The following points of interest have appeared.

1. Approximately one-half, forty-four per cent, of the students' study is carried on in the evening by artificial light.
2. Tuesday tends to be the day on which the most studying is done; Friday is the day when the least studying is done.
3. Sunday is a better study day than either Friday or Saturday, judged in terms of amount of time devoted.
4. The average daily study time for the entire group was one hour and fifty-one minutes. When those who did not study on a given day are eliminated this average rises to two hours and thirteen minutes.

---

[1] See May, Mark A.: Predicting Academic Success. *Journal of Educational Psychology*, Vol. XIV, 1923, pp. 429–440.

5. Towards the end of the week there is a pronounced tendency to shorten up on the duration of the study periods and a less pronounced tendency for fewer students to be studying.

6. Sixty-seven per cent of the students studied on Sunday, ninety-four per cent on Monday, over ninety-nine per cent on Tuesday, ninety-six per cent on Wednesday, ninety-seven per cent on Thursday, seventy-three per cent on Friday, and forty-one per cent on Saturday.

7. Sixty-three per cent of the students' study time was given to reading textbooks; nineteen per cent to outside readings; 8.5 per cent to working problems; and 9.5 per cent to writing themes and other papers. Eighty-two per cent of the total time was used in reading and eighteen per cent in writing.

8. Study time correlates +.317 ± .05 with scholastic standing, intelligence correlates +.565 ± .04 with scholastic standing; and intelligence correlates +.003 ± .06 with study time. With intelligence held constant, study time correlates +.382 ± .05 with scholastic achievement. Intelligence appears to be a better indicator of scholastic success than study time.

CONSTANT CHANGES IN THE STANFORD-BINET IQ

PSYCHE CATTELL

Harvard University

The Harvard University Growth Study, conducted by Professor Walter F. Dearborn, is now in its ninth year. Each year some one hundred fifty to two thousand of the children have been given an individual Stanford-Binet intelligence examination. By the end of the seventh year two or more Binet IQ's had been secured for one thousand one hundred eighty-three pupils. These one thousand one hundred eighty-three subjects were divided into groups according to the period of time that had elapsed between the first two Binet examinations: under three months, three to six months, six to twelve, twelve to eighteen, eighteen to twenty-four, twenty-four to thirty-six, thirty-six to forty-eight, forty-eight to sixty and sixty to seventy-two.[1]

The median amount by which the second IQ fell short of or exceeded the first is shown in Table I. It appears probable that the median gain of 5.0 points in the group repeating the test within a period of three months and of 3.8 when repeating it after a period of from three

TABLE I.—MEDIAN BINET IQ CHANGES AFTER VARYING INTERVALS OF TIME

Months between tests	No. of cases	Median differences	PE
0– 3	18	+5.0	±1.1
3– 6	54	+3.8	±0.8
6–12	308	+0.2	±0.4
12–18	174	+2.0	±0.6
18–24	64	−0.5	±0.9
24–36	92	−0.2	±0.9
36–48	203	−2.7	±0.4
48–60	51	+1.0	±1.5
60–72	320	−0.1	±0.4
0–72	1383	−0.02	±0.22

---

[1] It was found that the effect of practice did not carry over eighteen months, therefore a child that had been given several tests was used in more than one comparison provided that no test had been given between the two being compared or within a period of 18 months preceding the first. This procedure increased the number of comparisons by two hundred bringing the total up to one thousand three hundred eighty-three.

to six months resulted from the experience gained during the first testing. There, however, appears to be but little, if any, practice effect resulting from a test taken more than six months previous. The group of one hundred seventy-four cases with an interval of twelve to eighteen months made a median gain of 2.0 points, but the three hundred eight cases with an interval of 6 to 12 months showed a gain of only 0.2 of a point, from eighteen to twenty-four months there was a loss of 0 5 of a point

In an earlier study it was found that at the upper levels of intelligence certain group tests gave uniformly higher and others uniformly lower IQ's than the Stanford-Binet [1] The question arises as to whether this was caused by variations in the standards of one or both tests, by factors inherent in the test items, or since the group and individual tests were not always given in the same year, whether there was a real change in the relative intelligence level of the child.

When comparing the gains and loses of Stanford-Binet IQ's of bright and dull children, Garrison writes·

> Since there does seem to be a slight gain in the higher classes, it is evident that there is a slight practise effect, that the test is relatively easier in the higher ages, or that the IQ actually increases for the higher classes We feel that there are not enough data available yet to warrant definite conclusions [2]

Rugg and Carlton state that.

> Differences in retests will be approximately the same, irrespective of intelligence of the pupil.[3]

And Terman that

> . . the IQ remains almost equally constant for the three groups The greatest tendency to gain appears with the average group and the next greatest with the dull The differences, however are practically negligible It makes little difference whether the child was bright, average, or dull, how long an interval separated tests or what the age of the child was at the earlier test [4]

Each time-interval group was redivided into three IQ levels according to the average of the Binet IQ's below ninety, ninety to one hundred nine, one hundred ten and higher. The median of the differences of each group are plotted in Fig 1 The first bar of each group represents those with IQ's below ninety, the second those with IQ's from ninety to one hundred nine and third those of one hundred ten and above The amount by which the median child's second IQ fell short of or exceeded the first is given on the vertical scale. For example, in the second group in which the time between the two tests varied from six to twelve months, the median pupil with an IQ below ninety lost 1.4 points, the median of those of average ability gained 0.1 of a point and the median of those with IQ's of one hundred ten or higher gained 0.2. The same data are given in tabular form at the foot of the chart, the first row gives the median of the differences, the second the number of cases and the third the time interval between the two tests

---

[1] Cattell, Psyche Comparability of IQ's Obtained from Different Tests at Different IQ Levels *School and Society*, March 29, 1930

[2] Garrison, C S Additional Retests by means of the Stanford Revision of the Binet-Simon Tests *Journal Educational Psychology*, May, 1922, p. 311

[3] Rugg, Harold and Cecile Colloton Constancy of the Stanford-Binet IQ as Shown by Retests *Journal Educational Psychology*, Sept , 1921, p 320

[4] Terman, Lewis M "The Intelligence of School Children" Houghton Mifflin Co , 1919, pp 140 and 146

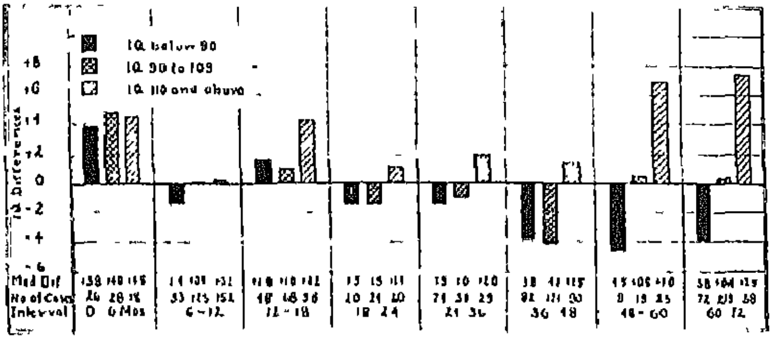


FIG. 1—Median changes in the Binet IQ of dull average and bright pupils after varying intervals of time.

When the tests were repeated after an interval of less than six months the median IQ of the pupils at all three levels of intelligence showed a gain, probably due, as stated above, to the effect of practice. The average group gained slightly more than either the bright or dull group, but the differences are too small to be of significance. As the time interval between the tests is increased beyond eighteen months a tendency for the low IQ's to decrease and the high IQ's to increase becomes evident.[1] The further apart the two tests the more marked is the tendency. When the last two groups are combined it is found that the median loss of the eighty-one cases with IQ's of below ninety was four points while the sixty-one cases with IQ's of one hundred ten or higher gained seven points The median change of those of average intelligence was only one-half of one point.

---

[1] When a child passed a total of five or more tests in years sixteen and eighteen, Professor Terman's correction was applied to the mental age (Genetic Studies of Genius, Vol. I, p. 42.) Since only a few of the children were over fourteen years of age at the time of the last test, the majority being under thirteen, the correction was needed in only a small proportion of the cases

The tendency for the low IQ's to lose and the high IQ's to gain is shown even more clearly in Fig 2. All these cases in which the interval

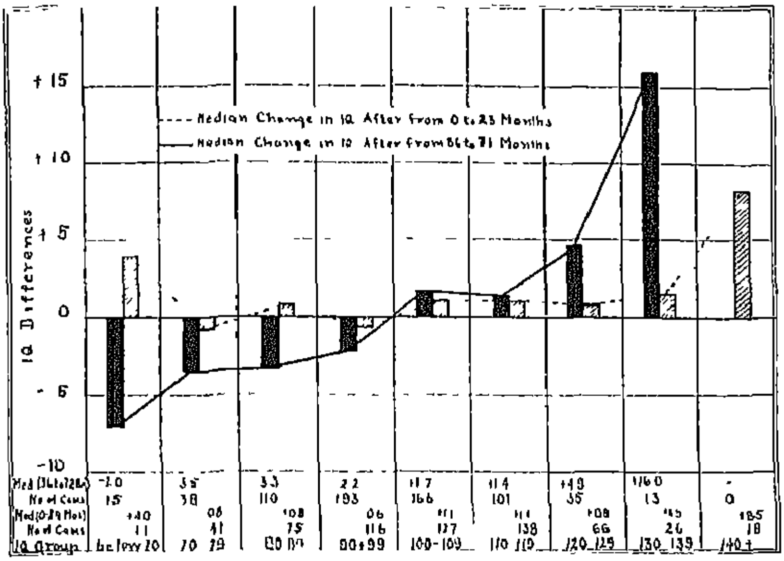


FIG 2 —Median Binet IQ changes at different intelligence levels

between the two Stanford-Binet tests was over three years were placed in one group and those with an interval of under two years in another Within each group the subjects were divided into ten point IQ groups, leaving those below seventy in one group and those above one hundred forty in another The medians of the differences between the first and second IQ of the pupils repeating the test after a period of from three to six years are represented by solid bars connected by a solid line. The tendency for the low IQ to become lower and the high IQ to become higher is marked The lowest group lost 7 5 points, the highest gained 16 0, while the medians of the changes of the remaining groups ranged between these two extremes

The median of the differences between the first and second IQ of the groups taking the two tests at intervals under two years gives a different picture. They are represented in shaded bars connected by a broken line in Fig. 2 No group had a median change of more than 1.5 points with the exception of the group below sixty IQ and the one above one hundred forty. There are included seventy-two that took the two tests within a period of six months The effect of practice carried over to the second test in these cases is probably sufficient to account for the excess of small gains over losses in the intermediate groups The lowest group only contains eleven cases and the gains of four points may well be due to chance errors. The gain of nearly nine points in the group of eighteen cases with IQ's of one hundred forty and above is not so easy to explain, it may also be due, at least in part, to chance errors, or it may be that the effect of practice is carried over a longer period in the case of exceptionally bright children than it is among other children (only three of the eighteen cases were retested within a period of less than six months). Finally it is possible, though it does not seem probable, that the relative intelligence level of the group was actually raised during the short time between the tests. More than one factor is probably involved.

There is, of course, a possibility that some of the changes in IQ are caused by variations in the testers Fifty-seven different examiners took part in the testing. All were instructors, assistants or graduate students in psychology or education at Harvard University. However, the general trend of the curves is too regular to be explained away by chance errors in the sample or by the personal equation of the examiners, it is more likely that such errors are responsible for the irregularities in the general trend.

The cause of the gain of the high IQ's and the loss of the low is not clear The factor of practice is ruled out by the fact that the greater the time interval between the tests, the greater the gain of the bright pupils, and the fact that the pupils of low intelligence lost. The gain in the high IQ's would be accounted for if the upper tests on the Binet scale were too easy, but the reverse has generally been supposed to be true and, in any case, this would not explain the loss of the low IQ's. It may be that the bright child has, or finds, more opportunities to acquire that type of information which aids in successfully passing the Binet tests than the older child of the same mental age. The greater interest in reading of the bright child and the lesser interest of the dull child may be a factor.

If there is a tendency for the range in IQ's to increase with age it would be sufficient to account for the median increase of the high IQ's and the decrease of the low IQ's. This seems to the writer the more probable *explanation* It is, of course, well known that the range of the distributions of heights and weights of children increase as the age increases, also intelligence when measured in terms of mental age. The theory of the constancy of the IQ is based on the assumption that though the intelligence of the bright child increases at a greater rate than the dull child, that of both the bright and the dull child remain constant in proportion to his chronological age. A child that has a mental age 20 per cent above his chronological age at school entrance would maintain the same proportional difference between his CA and MA throughout the period of growth Although the difference between his mental and physical age would increase, his IQ would remain constant.

The present findings indicate that those pupils that start school with IQ's above the average, not only increase in mental ability at a rate sufficient to keep the proportion between the life and mental age the same, but at a more rapid rate, causing the IQ to rise, while the dull child's mental age falls relatively to his CA, causing a drop in the IQ This is contrary to the findings of Terman and of Rugg and Carlton quoted above but is corroborated by Garrison who finds a slight increase in the higher classes and with Cyril Burt, Doll and others who have found a decrease in the IQ of sub-normal children and with Kuhlmann who writes that.

> *It has been known for some time that the IQ as found by tests that give* correct mental ages at all mental levels does not remain constant throughout *successive years, except for children with an initial IQ of 1 00 Initial intelligence* quotients below 1.00 tend to decrease, and above 1 00 they tend to increase as the *child grows older* [1]

Freeman in referring to sub-normal children makes the opposite statement·

> We may expect backward children to gain about as much from year to year and to continue to gain about as long as normal children or bright children.[2]

This would mean a marked increase in IQ of the backward children as age increases. It appears, however, from other statements made by Freeman that this is a slip of the pen and that what he had intended to write was probably that the dull child continued to grow as steadily and for as long a time as the normal and bright child, but at a slower rate. If this be the correct interpretation he is in agreement with the view held by Terman and his followers

---

[1] Kuhlmann, F. and Rose Anderson "The Forthcoming Revision of the Manual of Instruction for the Kuhlmann-Anderson Intelligence Tests" The Educational Test Bureau, Minneapolis, Minn, p 118.

[2] Freeman, Frank N. "Mental Tests" *Houghton Mifflin*, p 357

CONCLUSIONS

1. When a Stanford-Binet test is repeated within a period of three or four months the experience gained while taking the first test appears to result in a median gain of four or five points in the second IQ No significant difference in amount of gain made by bright, average or dull children was found The practice effect carried beyond six months appears to be insignificant.

2. A definite tendency was found for the pupils of high intelligence to gain and for those of low intelligence to lose in IQ as they become older.

SOME RELATIONSHIPS BETWEEN ALGEBRA AND GEOMETRY

DORRIS MAY LEE

Glendale City Schools, Glendale, Calif

AND

J MURRAY LEE

Director of Research, Burbank, Calif

1. What is the relationship between ability in algebra and geometry?

2. What is the relationship between achievement in algebra and geometry?

3. What percentages of pupils show differences between algebra and geometry, in respect to ability and achievement?

4. Are the students who continue with geometry a more select group than all the pupils who took algebra?

5. Do students tend to get better or poorer grades in geometry than in algebra?

An attempt will be made to answer the above questions using data made available from two previous studies Since these studies were conducted a year apart, and two of the schools were used in both studies, there were complete records in both algebra and geometry of one hundred eighty-one pupils These records consist of scores on the Lee Test of Algebraic Ability[1] and the Lee Test of Geometric Aptitude[2] given before the pupils had begun the study of the respective subjects, scores on an algebraic achievement test[3] and on the Renfrow Geometry Test[4] given at the end of the first semester of each subject, and the first semester's mark in both algebra and geometry

There has not been an attempt to follow-up each case that took algebra. There was loss through failures, drop-outs, and moving Whether the pupils that moved would be selected from the upper or lower ranges of ability is not known, but the writers have assumed that the factor of moving has not caused a distortion of the data

---

[1] Published by Public School Publishing Co , Bloomington, Ill , 1930

[2] Published by Southern California School Book Depository, Los Angeles, Calif , 1931

[3] A non-standardized test of 60 items

[4] Published by C A Gregory Co , Cincinnati, Ohio

What is the relationship between ability in algebra and in geometry? There is one published study, A. L. Rogers,[1] that furnishes data that is comparable with the data in the present study. Dr. Rogers found the correlation between algebraic and geometric ability to be .52 for one school and .38 for another with an average correlation of .47 ± .03. When this average was corrected for attenuation it became .54 ± .02.

Correlations were found between the Lee Test of Algebraic Ability and the Lee Test of Geometric Aptitude for the one hundred eighty-one pupils whose complete records were available. These correlations with the corrections for attenuation are given in Table I.

TABLE I.—THE CORRELATION BETWEEN ABILITY IN ALGEBRA AND ABILITY IN GEOMETRY AS SHOWN BY THE CORRELATION BETWEEN THE LEE TEST OF ALGEBRAIC ABILITY AND THE LEE TEST OF GEOMETRIC APTITUDE

	r(T.A.A.)(T.G.A.)	r(T.A.A.)(T.G.A.) corrected[1]
School *A*	.55 ± .04	.60 ± .04[2]
School *B*	.57 ± .07	.62 ± .07

[1] Kelley, Truman L.: "Statistical Method." New York: Macmillan Co., 1923, p. 204, Formula 165a.

[2] *Ibid*: P. 209, Formula 164.

From the results of these two studies the conclusion can be drawn that the correlation between ability to do algebra and ability to do geometry probably lies between .50 and .65.

What is the relationship between achievement in algebra and achievement in geometry? Most of the studies in this field have dealt with school marks. The correlation between the final mark in algebra and the final mark in geometry has been found. These studies are summarized in Table II.

In the present study there are two measures of achievement, the marks for the first semester of algebra and geometry and the scores on achievement tests given at the end of the first semester of the respective subjects. These correlations are presented in Table III.

It is interesting to note that in neither school is the correlation as high between achievement tests as it is between marks. *This would*

---

[1] Rogers, A. L.: Experimental Tests of Mathematical Ability and Their Prognostic Value. *Teachers College Contributions to Education, No. 89*. New York: Teachers College, Columbia University, 1918, pp. 79-80.

seem to indicate an additional factor, separate from achievement, that is influencing the marks

From a study of Tables II and III, the correlation between achievement in algebra and geometry probably lies between 40 and 70.

TABLE II.—CORRELATIONS BETWEEN ACHIEVEMENT IN ALGEBRA AND ACHIEVEMENT IN GEOMETRY, AS MEASURED BY TEACHERS' MARKS, FROM PREVIOUS STUDIES

Author	No of cases	Correlations
Burris[1]	16 schools	57
Crathorne[2]	1000 (approx)	52 ± 02
Winegardner[3]	417	509 ± 024

[1] Burris, W P. Correlations of the Abilities Involved in Secondary School Work, Edited by E L Thorndike *Columbia University Contributions to Education*, Vol. XI, No 2, 1003

[2] Crathorne, A. R "The Theory of Correlation Applied to School Grades," The Reorganization of Mathematics in Secondary Education, Chap X The Mathematical Association of America, 1923

[3] Winegardner, J H The Relation of Success in Mathematics to Success in Physics and Chemistry in High School *Unpublished Master's thesis*, Department of Education, Stanford University, 1929

This fact and Table I tends to show that the correlation between ability in algebra and geometry is usually higher than that of achievement, and further, that the correlations between ability are more consistent than those between achievement

TABLE III —CORRELATIONS BETWEEN ACHIEVEMENT IN ALGEBRA AND ACHIEVEMENT IN GEOMETRY, AS MEASURED BY TEACHERS' MARKS AND ACHIEVEMENT TESTS

	N	*r* between marks	*r* between tests
School *A*	136	474 ± 044	410 ± 048
School *B*	45	709 ± 050	566 ± 069

What percentages of pupils show differences between algebra and geometry, in respect to ability and achievement? An attempt to answer this question is another means of attacking the problem of the relationship between these two subjects The method of finding such differences as called for by the question has been determined by Kelley [1]

---

[1] Kelley, Truman L · A New Method for Determining the Significance of Differences in Intelligence and Achievement Scores *Journal of Educational Psychology*, Vol XIV, September, 1923

The method used is to find the value of $\sigma_{d_\infty}/\sigma_d$, then by referring to Table IV in Kelley's article, change this ratio to a percentage. This then gives the percentage of pupils that show differences between algebra and geometry that are more frequent than the chance factors in the measures would produce.

The data and formulae which are needed in this evaluation are as follows:

$$\sigma_{d_\infty} = \sqrt{2 - r_{1I} - r_{2II}} \qquad \text{(see Kelley}^1\text{)}$$

and

$$\sigma_d = \sqrt{2 - 2r_{12}}$$

Table IV.—Reliability Coefficients of the Various Tests Which Were Used

Tests	Reliability Coefficients
Test of Algebraic Ability	93
Test of Geometric Aptitude	91
Algebra Achievement Test	.95
Geometry Achievement Test	85

The other correlation coefficients needed are found in Tables I and III.

The percentage of differences in individual test scores, as shown by the application of the above formulæ and Table IV (Kelley[1]) between algebra and geometry in respect to achievement is forty-one per cent for School A and thirty-four per cent for School B and in respect to ability is forty per cent for School A and thirty-nine per cent for School B. This indicates that there is about as large a percentage of differences found between achievement and also ability tests in algebra and geometry as there is between such tests in the Stanford Achievement Battery, as Arithmetic Computation and Paragraph Meaning.[1]

Using the same technique, differences between the marks in algebra and in geometry were determined. Using .70 as a reliability coefficient for marks and the correlations between marks as given in Table III, the differences were found to be thirteen per cent and zero per cent. If higher reliabilities of marks had been assumed these differences would have been larger, if however, a lower coefficient had been used the differences would have been even less. These differences seem to further indicate there is another factor entering into the marks in both algebra and geometry than those that can be accounted for by achievement or ability.

---

[1] *Ibid.*

Are the students who take geometry a more select group than those who take algebra? In answering this question three sets of data will be scrutinized. These are the algebra marks, algebra achievement scores, and the scores on the Lee Test of Algebraic Ability. The records of those pupils for whom both algebra and geometry scores are available will be compared with the algebra records of all the pupils.

First, the algebra marks of the pupils continuing with geometry are compared to the algebra marks given to the whole group. These facts are given in Table V.

TABLE V.—A COMPARISON OF THE MARKS RECEIVED IN FIRST SEMESTER ALGEBRA BY THOSE PUPILS CONTINUING IN GEOMETRY, WITH THOSE RECEIVED BY ALL PUPILS

Algebra marks	Number receiving each mark		Percentage of pupils continuing
	All	Continuing	
A	34	27	79.4
B	125	85	68.0
C	152	65	42.8
D	12	3	25.0
F	71	1	1.4
Total	394	181	45.9

In studying Table V, the fact that pupils may have moved, or postponed taking algebra should be considered. The last column shows that 79.4 per cent of the pupils who received "*A*"s in algebra continued with geometry. From this it drops to 68.0 per cent of the "*B*"s continuing. For the "*C*"s it drops to 42.8 per cent then to 25 per cent of the "*D*"s and to 1.4 per cent of the "*F*"s or failures. It is obvious that a much greater percentage of the pupils receiving "*A*"s and "*B*"s continue than do those receiving "*C*"s and "*D*"s. Most students receiving "*D*"s and "*F*"s in algebra consider that they have had all the experience with mathematics courses that they intend to have and therefore do not take geometry.

This condition would seem to indicate that there is a need for careful guidance of the pupils who are to take algebra and the provision of some other course in mathematics for those pupils who will probably fail or receive "*D*"s. This course should be much easier and cover a much broader mathematical field than the regular algebra course

Second, the achievement scores of all pupils are compared with those of the pupils who continue with geometry The records of the two groups are given in Table VI

TABLE VI.—A COMPARISON OF THE SCORES RECEIVED ON AN ALGEBRAIC ACHIEVEMENT TEST GIVEN AT THE END OF THE FIRST SEMESTER, BY ALL THE PUPILS WITH THOSE OF THE PUPILS WHO CONTINUED WITH GEOMETRY

ALL PUPILS		CONTINUING PUPILS	
M_1 =	25.44	M_2 =	29.32
σ_1 =	9.20	σ_2 =	7.58
σ_{M_1} =	.46	σ_{M_2} =	.56
N =	394	N =	181

It can be seen from an inspection of Table VI that the mean score of the continuing group is 3.88 points higher than the mean made by all the pupils Does this indicate that there is a real difference between these two groups or is it merely a difference that might be caused by chance? This question can be answered by finding the ratio of the observed difference to the standard deviation of that difference $\left(\frac{M_2 - M_1}{\sigma_{dM}}\right)$. If this ratio is over three, it is fairly certain that there is an actual difference, not a chance difference The formula used to find the sigma of the differences of the means is:

$$\sigma_{dM} = \sqrt{\sigma^2_{M_1} + \sigma_{M_2}^2 - 2r\sigma_{M_1}\sigma_{M_2}}$$

where σ_{M_1} is the σ of the mean of all pupils (see Table VI)
and σ_{M_2} is the σ of the mean of the continuing pupils.

In evaluating the above formula if r is given a value of the σ_{dM} value will be a maximum and $\frac{M_2 - M_1}{\sigma_{dM}}$ is a minimum

Evaluating

$$\sigma_{dM} = .73 \text{ and } M_2 - M_1 = 3.88,$$

hence

$$\frac{M_2 - M_1}{\sigma_{dM}} = 5.35$$

Since the ratio of the observed difference to the standard deviation of that difference is 5.35, it is certain that the difference is real as far as achievement in algebra is concerned and that the pupils continuing with geometry were superior in achievement to the group as a whole

Third, the scores of all the pupils on the Lee Test of Algebraic Ability are compared with those of the pupils continuing with geometry. The same method of comparison is used for the ability scores as is used for the achievement scores. The records of the two groups on the ability test are given in Table VII.

TABLE VII.—A COMPARISON OF THE SCORES RECEIVED ON THE LEE TEST OF ALGEBRAIC ABILITY BY THOSE PUPILS CONTINUING IN GEOMETRY WITH THOSE OF ALL PUPILS

ALL PUPILS	CONTINUING PUPILS
$M_1 = 81.63$	$M_2 = 92.94$
$\sigma_1 = 24.63$	$\sigma_2 = 21.81$
$\sigma_{M_1} = 1.24$	$\sigma_{M_2} = 1.62$
$N = 394$	$N = 181$

It can be seen from Table VII, that the mean score of the group that continued with geometry is 11.32 points higher than the mean of all the pupils. Evaluating this difference—

$$\text{Assuming } r = 0, \text{ then } \sigma_{dM} = 2.04, \text{ and } \frac{M_2 - M_1}{\sigma_{dM}} = 5.55.$$

Since this ratio is equal to 5.55 it is certain that the difference was real as far as ability to do algebra, as measured by the Lee Test of Algebraic Ability is concerned and that pupils continuing with geometry were superior in algebraic ability to the group as a whole.

The three sets of data, marks and achievement record in first semester algebra and the scores on the ability test all tend to prove that the group that continued with geometry were a select group as compared with the total group that took algebra. The poorer students, academically speaking, do not continue with their mathematics. If by careful educational guidance we would place the pupil where he belonged in the first place, instead of letting him find it by trial and error, what a large amount of time, money, and discouragement could be saved.

Do students tend to get better or poorer grades in geometry than in algebra? One means of answering this question is from an inspection of the scattergrams between marks in algebra and in geometry.

Chart 1 shows that pupils tended to get poorer grades in geometry more often than they got better ones as compared with their grade in algebra.

It is interesting to note that only one pupil raised his mark as much as two points. This pupil in School A received a C in algebra and an

A in geometry. His ability test in algebra showed he was quite low but in geometry his ability score went up to the 75th percentile of the group His achievement tests agreed with the marks, in fact, all records indicated that he was only low average in algebra and superior in geometry. He is the best example of marked differences favoring geometry that is in the study.

CHART I.—A COMPARISON BETWEEN MARKS RECEIVED IN GEOMETRY WITH THOSE RECEIVED IN ALGEBRA

School A

Algebra	Geometry *F*	*D*	*C*	*B*	*A*	
A		1	1	9	**7**	18
B	2	8	17	**25**	11	63
C	9	13	**18**	13	1	54
D			1			1
F						
	11	23	36	47	19	136

School B

Algebra	Geometry *F*	*D*	*C*	*B*	*A*	
A			1	4	**4**	9
B		3	14	**3**	2	22
C	2	5	**3**	1		11
D		**2**				2
F	**1**					1
	3	10	18	8	6	45

The figures in bold face type indicate the pupils that received the same mark in both subjects Those above this diagonal line received a lower mark while those below it received a higher mark. In School A, fifty pupils received the same mark, twenty-six improved their mark while sixty lowered it. This situation is to be expected since in the case of algebra these pupils were a select group in the upper ranges, but in geometry these same cases were redistributed over a normal distribution. It would then be expected that the chances for a pupil to get a lower mark in geometry are greater than his chances to get a higher mark In School A, this chance of getting a lower mark was a little over two to one, while in School B it was nearly ten to one

Another comparison between algebra and geometry that is worth noting is the differences in the percentage of each mark given. Taking the two school systems that were used in both studies, percentages of each mark given are compared in Table VIII.

There is a slightly greater per cent of "*A*"s given in geometry, however the outstanding difference is in the percentage of "*D*"s This might be accounted for by the fact that algebra is much more definite and the failing point can be determined with more ease than in the case of geometry.

TABLE VIII.—A COMPARISON OF THE NUMBER AND PERCENTAGE OF EACH MARK GIVEN IN ALGEBRA AND GEOMETRY

Mark	Algebra		Geometry	
	Number	Per cent	Number	Per cent
A	34	8.6	38	12.2
B	125	31.7	98	31.4
C	152	38.6	90	28.9
D	12	3.0	51	16.3
F	71	18.0	35	11.2
Total	394	99.9	312	100.0

SUMMARY

Comparisons made between algebra and geometry using marks, achievement test scores, and ability test scores have led to the following conclusions.

1. The relation between ability to do algebra and ability to do geometry as expressed by correlation coefficients, probably lies between .50 and .65

2. The correlation between achievement in algebra and geometry probably lies between .40 and .70

3. The correlations between ability in algebra and geometry are usually higher and more consistent than those of achievement.

4. About 40 per cent of the pupils show differences between algebra and geometry in respect to both ability and achievement that can not be attributed to chance

5. Some factor other than ability and achievement is entering into school marks in these two subjects

6. Students who take geometry are a select part of the group that took algebra the year before

7. Pupils receiving low marks in algebra do not as a rule take geometry

8. Pupils have a greater chance in geometry of getting a mark lower, rather than higher, than they got in algebra.

9. It seems that it is more difficult to determine the failing point for a pupil in geometry than it is in algebra.

The educational implications of the above findings seem to be

1. Better guidance of a pupil in geometry is possible by using tests of ability to do geometry than by using the algebra record. The two however should be used together.

2. More careful educational guidance should be given pupils who are ready to enroll in algebra, and the probable failures eliminated. A course should be provided for these pupils which is much easier and covers as broad a mathematical field as possible

3. Marks should be given on the basis of achievement. In order to eliminate other factors which are now determining the mark, such objective measures of achievement as are now obtainable should be used. These objective tests should be given several times throughout the semester to give as many measures as is possible

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII November, 1931 Number 8

SCHOOL ACHIEVEMENT IN RELATION TO MENTAL AGE—A COMPARATIVE STUDY*

ANDREW W BROWN AND CHRISTINE LIND

Institute for Juvenile Research (Chicago, Ill.)

I. STATEMENT OF PROBLEM

It is fairly well established that in the regular public school where the children are of average intelligence, the achievement of those with intelligence quotients below one hundred is, in general, above that which would be expected of their mental age, while the achievement of the children with intelligence quotients above one hundred is likely to be below their mental age and the further the intelligence quotient is below or above one hundred the greater the discrepancy between the mental age and achievement. In other words, the accomplishment ratio (AR) which is the ratio between the educational age and the mental age correlates negatively with the intelligence quotient. This is the case in the public school where the majority of the children are of average intelligence.

But is it a general characteristic of all dull children that their achievement be above their mental age and of all bright ones that their achievement be below their mental age, or is this only relative to the group in which the children happen to be placed? To be more specific. It is known that children in the regular public school with intelligence quotients from seventy to ninety, have, in general, achievement ratings above their mental age, but if these same children were in a group where they were relatively bright, would their achievement ratings still be above their mental age or would it fall below? Is the difference between the mental age and achievement age the same for each school subject or do some subjects show a greater variation than others? And again, at what level of intelligence is the variation the greatest? It is these problems that this study attempts to investigate.

---

* Studies from the Institute for Juvenile Research, Chicago, Series C, 165.

II. METHOD OF STUDY

Two groups of children were selected for study, one group from the academic department of the Lincoln State School for the feeble-minded and one from the Glenwood Manual Training School.

At Lincoln only those children are admitted to the academic department who have intelligence quotients above 50. There were four hundred children in the academic department but only those above grade two were used for this study. Shortly after the children are committed to the institution, they are given the Stanford-Binet examination by the resident psychologist. It is the mental age and intelligence quotient secured by this examination that is used here. If the mental age had been secured more than a month prior to the time of the study, it was brought up to date by multiplying the intelligence quotient by the present chronological age.

Glenwood is a private school for dependent boys. There is no selective factor other than dependency in the selection of the students. The average intelligence rating of the entire school population is around .95. The mental age and intelligence quotient used in this study were secured from the Haggerty Intelligence Examination Delta 2.

The children of both groups were given the following tests of the Stanford-Achievement Series:

1. Paragraph Reading.
2. Word Reading
3. Arithmetic Reasoning.
4. Arithmetic Computation.
5. Spelling.

In addition to the above, samples of writing were secured from the Lincoln Group and scored according to the Thorndike Writing Scale. One hundred twenty-five of this group were given the Goodenough Drawing Test.[2]

---

[1] Terman, L. M., G. M. Ruch, and T. L. Kelley "Manual of Directions for Stanford-Achievement Tests." World Book Company

[2] Goodenough, Florence: "Measurement of Intelligence by Drawings" World Book Company.

All of these tests except writing have been standardized for various age levels, and if there were a perfect relation (assuming the tests are equally well standardized) these age levels for the various abilities should presumably be the same Therefore, the amount of difference between the achievement ages and the mental age is a measure of the degree of relationship

III. PRESENTATION OF DATA

(*A*) *Retarded Group* (*Lincoln*).—Table I shows the difference for the Lincoln group between the mental age and the ages achieved on the other tests for the three levels of intelligence—fifty to fifty-nine IQ, sixty to sixty-nine IQ, seventy to seventy-nine IQ, for each of the different subjects· Reading paragraphs (RP), reading words (RW), arithmetic reasoning (AR), arithmetic computation (AC), spelling (Sp), writing (Wr), and drawing (Dr). The table shows that in reading paragraphs, the sixty-five children from fifty to fifty-nine IQ made a total of five hundred fifty-seven months above their mental age, and two hundred twenty-nine below their mental age. The average algebraic difference [(557–229) ÷ 65)] of 5 04 months indicates a tendency for this group, in general, to make a rating higher than their mental age in this subject. The eighty-two children with IQ's between sixty and sixty-nine make a total of five hundred eighty months above their mental age and four hundred ninety-seven months below their mental age On the average, they are 1.01 months above their mental age in reading paragraphs. The thirty-six children with IQ's from seventy to seventy-nine make a total of ninety-four months above and three hundred thirty-two months below their mental age, with an average of 11.5 months below their mental age The differences between the mental age and the educational age for each of the other subjects are read in the same way The significant facts of this table are presented in Fig 1

Figure 1 shows the variation from the mental age of the various abilities at the different levels of intelligence. The mental age is the standard from which the variation is measured and is indicated by the zero line. The difference in months between the mental age and the ages in the different subjects is represented on the horizontal axis. For example. The children with IQ's from fifty to fifty-nine are, on the average, 5 04 months higher in paragraph reading (RP) than their mental age The children with IQ's from sixty to sixty-nine are 1.01 months higher, and those with IQ's from seventy to seventy-

TABLE I.—LINCOLN GROUP SHOWING THE DIFFERENCES IN MONTHS BETWEEN THE ACHIEVEMENT AGE IN SCHOOL SUBJECTS AND THE MENTAL AGE FOR THE VARIOUS LEVELS OF INTELLIGENCE

	Fifty to fifty-nine IQ				Sixty to sixty-nine IQ				Seventy to seventy-nine IQ			
	Number of cases	Total differences in months		Average algebraic difference	Number of cases	Total differences in months		Average algebraic difference	Number of cases	Total differences in months		Average algebraic difference
		Plus	Minus			Plus	Minus			Plus	Minus	
Reading paragraphs	65	557	229	+ 5 04	52	550	497	+ 1 01	36	91	332	− 6 61
R. W.	65	825	177	+ 9 97	52	800	228	+ 6 98	36	168	412	− 6 77
Arithmetic reasoning	66	542	283	+ 3 92	52	396	439	+ 0 52	38	69	367	− 7 85
Arithmetic computation	66	658	162	+ 7 51	52	455	305	+ 1 79	38	46	444	−10 47
Spelling	66	53	822	−11 65	52	71	1551	−18 05	38	22	1119	−28 87
Writing	66	550	459	+ 1 38	52	427	536	− 1 33	38	66	536	−13 95
Drawing	34	300	363	− 1 85	58	571	840	− 4 79	25	70	602	−22 21

nine are, on the average, 6.61 months below their mental age in this subject.

Tables II, III, and IV give the distribution of the differences in months between the mental ages and the achievement ages in each of the school subjects for each level of intelligence Most of the cases in the fifty to fifty-nine IQ group lie within a range from positive forty months to negative forty months. Drawing shows the greatest variation.

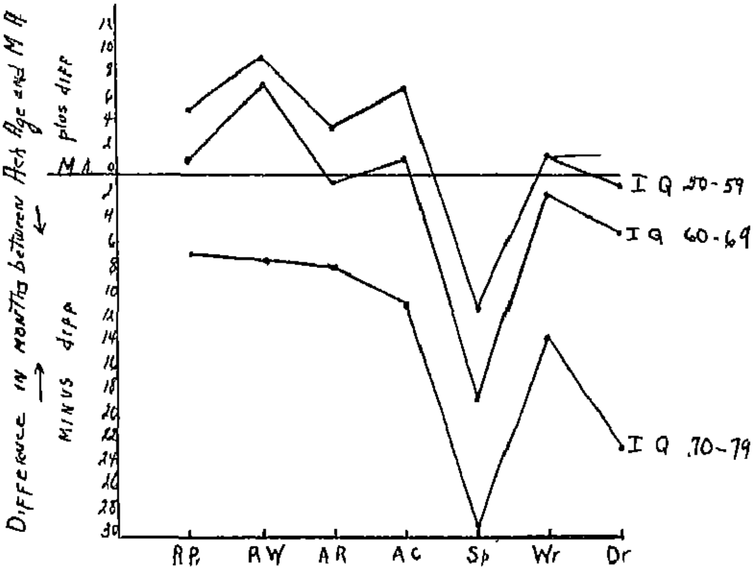


Fig. 1 —Lincoln Group showing the differences in months between the mental age and the achievement ages

In the sixty to sixty-nine IQ group most of the cases lie within a range from positive thirty months to negative fifty months In this group also the achievement in drawing has the widest range extending from positive seventy-five months to negative sixty-five months These two groups have a range of about eighty months. The seventy to seventy-nine IQ group has a smaller range, most of the cases lying between positive twenty months and negative forty months, or a range of sixty months. These tables show that the range of the cases for each subject in the three groups is about the same. In each of the groups, drawing has the widest range. Since all of the ranges are about the same, the distribution of the cases seems to show that each group has about the same scatter and that the significant facts lie in the change of the average algebraic difference rather than in the range of distribution. The average algebraic differences have been shown in Fig 1.

TABLE II.—FIFTY TO FIFTY-NINE IQ LINCOLN GROUP SHOWING THE DISTRIBUTION OF DIFFERENCES IN MONTHS BETWEEN MENTAL AGE AND ACHIEVEMENT AGES

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling	Writing	Drawing
80–84		1					
75–79							
70–74							
65–69						1	1
60–64							
55–59						1	1
50–54							
45–49							
40–44						1	
35–39	2		1				1
30–34	1	3		3		1	
25–29	2	5	1	2		5	
20–24	5	7	5	7		3	3
15–19	3	8	8	5		2	2
10–14	12	13	12	11	1	7	1
5– 9	10	9	8	16	4	5	2
0– 4	11	6	12	8	9	6	3
5– 1	5	3	6	9	7	9	5
10– 6	6	5	4	3	10	7	4
15–11	4			1	11	3	2
20–16		2	4		7	9	2
25–21	2	2	1	2	11	7	2
30–26	1		1	1	3	2	1
35–31		1	1	1			
40–36	1		2		2		1
45–41					1		1
50–46							1
55–51							1

Figure 1 shows that, in general, the lower the intelligence the higher the achievement in relation to the mental age. The children with IQ's between fifty to fifty-nine are doing work decidedly above their mental age level in all the subjects except spelling and drawing.

Those with IQ's between sixty and sixty-nine more nearly approach their mental age in all subjects except spelling and drawing, while those with IQ's from seventy to seventy-nine have achievement levels decidedly below their mental age in all subjects.

(*B*) *Average Group* (*Glenwood*) —The Glenwood Group was selected in order to determine whether the tendency for the brightest retarded children to make achievement ratings below that which would be expected for their mental age, and the dullest ones to have achieve-

TABLE III —SIXTY TO SIXTY-NINE IQ LINCOLN GROUP SHOWING THE DISTRIBUTION OF DIFFERENCE IN MONTHS BETWEEN MENTAL AGE AND ACHIEVEMENT AGES

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling	Writing	Drawing
70–74							1
65–69							
60–64							1
55–59							1
50–54				1			
45–49							
40–44		1					1
35–39							
30–34		1		2		1	2
25–29	3	0	1	1		1	3
20–24	4	0	3		1	4	3
15–19	6	9	7	3	1	0	3
10–14	16	15	6	8	2	5	4
5– 9	13	11	11	17	1	7	3
0– 4	9	11	17	16	1	16	7
5– 1	11	7	10	14	8	11	1
10– 6	5	7	12	11	9	9	4
15–11	2	4	7	6	7	9	1
20–16	2	1	3		20	4	5
25–21	3		2	2	11	6	1
30–26	3	2			9	2	1
35–31	1		1		3	1	4
40–36	2	1	1		4		4
45–41	1				2		
50–46	1				2		3
55–51							1
60–56			1	1	1		2
65–61							1

ment ratings above their mental age, is typical of the retarded children, or whether the same tendency is present in children who are not retarded.

The Glenwood group represents a fairly average group of boys The range of intelligence quotients was from fifty-five to about one hundred forty. The average intelligence quotient of the entire school population is about ninety-five.

TABLE IV - SEVENTY TO SEVENTY-NINE IQ LINCOLN GROUP SHOWING THE DISTRIBUTION OF DIFFERENCE IN MONTHS BETWEEN MENTAL AGE AND ACHIEVEMENT AGES

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling	Writing	Drawing
50–54							
45–49	.	1					
40–44							
35–39							
30–34							
25–29		.					1
20–24	.	1			1	1	
15–19	2	2					
10–14	1	3	4	1		2	3
5– 9	5	3	2	4		1	
0– 4	4	5	7	4		7	1
5– 1	7	5	3	4	1	2	2
10– 6	7	4	8	4		5	4
15–11	3	2	4	9	3	5	1
20–16	1	3	3	4	6	4	2
25–21	2		3	6	5	3	
30–26	.	2	2	1	8	2	4
35–31	1	1	1		2	2	2
40–36	2	1	1	1	4		2
45–41	.	1			2	2	2
50–46					1		
55–51	.		.	.	3	1	
60–56		2			1	1	
65–61	1				1		2
70–66	.						1
75–71							
80–76	. .			.		.	1
85–81							

The group was divided into six IQ levels: seventy-nine and below, eighty to eighty-nine, ninety to ninety-nine, one hundred to one hundred nine, one hundred ten to one hundred nineteen, one hundred twenty and above. The difference between the mental ages and the achievement ages for the various subjects was computed and tabulated in the same way as for the Lincoln group.

Table V presents the same data for the Glenwood group which Table I presents for the Lincoln group. This table shows that the

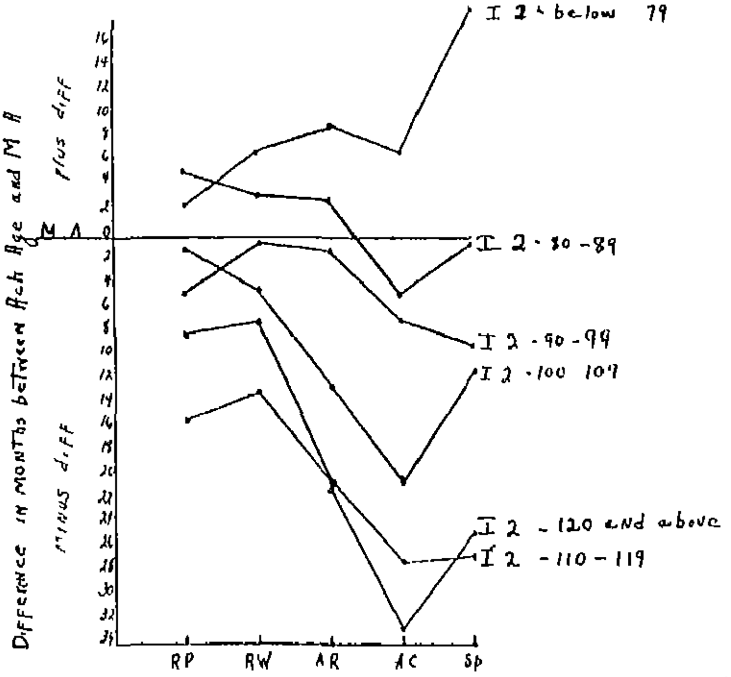


FIG 2.—Glenwood Group showing the difference in months between the achievement age in subjects and the mental age for the various levels of intelligence

twenty-seven children with IQ's of seventy-nine and below, made a total of one hundred thirty-eight months above their mental age and eighty-one months below their mental age in reading paragraphs. The algebraic difference $((138 - 81) \div 27)$ of 2.11 months indicates a tendency for this level to make a rating higher than their mental age in this subject. The thirty-one boys with IQ's from eighty to eighty-nine made a total of two hundred fifty-nine months above their mental age and one hundred eleven months below their mental age in reading

TABLE V.—GLENWOOD GROUP SHOWING THE DIFFERENCE IN MONTHS BETWEEN THE ACHIEVEMENT AGE OF SCHOOL SUBJECTS AND THE MENTAL AGE FOR THE VARIOUS LEVELS OF INTELLIGENCE

	Below seventy-nine IQ				Eighty to eighty-nine IQ				Ninety to ninety-nine IQ			
	Number of cases	Total differences in months		Average algebraic difference	Number of cases	Total differences in months		Average algebraic difference	Number of cases	Total differences in months		Average algebraic difference
		Plus	Minus			Plus	Minus			Plus	Minus	
Reading paragraphs	27	138	51	+ 2.11	31	259	111	+ 4.77	41	122	318	− 4.73
Reading words	27	232	58	+ 6.44	31	222	106	+ 3.74	41	157	173	− 0.39
Arithmetic reasoning	27	264	44	− 8.85	31	245	138	+ 3.54	40	172	225	− 1.40
Arithmetic computation	27	253	77	+ 6.51	31	129	303	− 5.61	41	120	412	− 7.12
Spelling	27	574	134	+16.29	31	171	199	− 0.90	41	50	449	− 9.00

	One hundred to one hundred nine IQ				One hundred ten to one hundred nineteen IQ				One hundred twenty and above			
	Number of cases	*Total differences in months*		Average algebraic difference	Number of cases	*Total differences in months*		Average algebraic difference	Number of cases	Total differences in months		Average algebraic difference
		Plus	Minus			Plus	Minus			Plus	Minus	
Reading paragraphs	31	116	192	− 1.03	15	43	165	− 8.13	16	21	266	−15.31
Reading words	31	116	252	− 4.38	15	23	133	− 7.33	16	45	251	−12.91
Arithmetic reasoning	*31*	*55*	*450*	*−12.61*	*15*	*0*	*303*	*−20.20*	*15*	*36*	*341*	*−20.33*
Arithmetic computations	30	0	627	−20.60	15		407	−27.13	16	24	550	−32.91
Spelling	31	90	457	−11.86	15		406	−27.06	15	8	389	−25.06

TABLE VI—SEVENTY-NINE IQ AND BELOW

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling
40–44					
35–39					
30–34	1				1
25–29	1	3	2	1	2
20–24		1	2	2	1
15–19	1	2	5	5	4
10–14	2	3	6	6	1
5– 9	6	8	5	2	2
0– 4	3	6	1	2	0
5– 1	7		3	2	2
10– 6	3	1	1	4	3
15–11	3	1	1	2	2
20–16		1	1	1	1
25–21		1			1
30–26					
35–31					
40–36					1

TABLE VII—EIGHTY TO EIGHTY-NINE IQ

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling
40–44					
35–39					
30–34	1	1	1		
25–29	1	1			
20–24	2	3	4	1	2
15–19	4	3	3	2	1
10–14	4	2	3	2	8
5– 9	4	1	6	5	
0– 4	5	8	2	5	6
5– 1	2	5	2	1	2
10– 6	3	2	6	3	3
15–11	3	4	1	5	4
20–16	1		1	2	1
25–21	1	1	1		2
30–26				1	
35–31			1	2	
40–36				2	1

paragraphs. The average algebraic difference for this level was 4.77 months above their mental age. The forty-one boys with IQ's from ninety to ninety-nine had an average of 4.73 months below their mental age in reading paragraphs. The differences between the mental age and subject ages for all the levels are read the same way. The significant facts of this table are presented in Fig. 2, this figure presents data for the Glenwood group similar to that which Fig. 1 presents for the Lincoln group.

TABLE VIII - NINETY TO NINETY-NINE IQ

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling
40-44					
35-39				1	
30-34			1		
25-29					
20-24	1	2		1	
15-19	1	2	4	2	
10-14	4	2	3	1	3
5- 9	4	8	5		7
0- 4	1	6	2	5	1
5- 1	11	7	9	9	8
10- 6	6	8	7	7	4
15-11	8	3	7	3	4
20-16	1	3		5	5
25-21	1			3	4
30-26	3		1	3	2
35-31					2
40-36					1
45-41					
50-46				1	
55-51					
60-56			1		
65-61					
70-66					
75-71					

Figure 2 shows the variation from the mental age of the various subjects at the different levels of intelligence. The mental age is the standard from which the variation is measured and is indicated by the zero line. The difference in months between the mental age and the ages in the different subjects is indicated on the vertical axis. The different subjects are indicated on the horizontal axis

Tables VI to XI show the distribution of differences in months between the mental ages and the achievement ages in each of the school subjects for each of the IQ levels. In this group most of the *cases fall within a range of sixty months. As in the Lincoln group* the range of the cases for each subject is about the same. The significant facts seem to lie in the change of the average algebraic differences rather than in the range of distribution.

TABLE IX.—ONE HUNDRED TO ONE HUNDRED NINE IQ

	Reading paragraphs	Reading words	*Arithmetic* reasoning	*Arithmetic* computation	Spelling
40–44	1				
35–39					
30–34	1				
25–29	1	1			
20–24					1
15–19	2	1			
10–14	2	3	1		2
5– 9	1	4	7	1	2
0– 4	1	3		2	3
5– 1	10	4	2	1	2
10– 6	6	2	3	5	1
15–11	2	0	3	5	5
20–16	2	3	4	3	4
25–21		3	4	3	4
30–26	2	1	3	4	5
35–31			3	3	1
40–36			1		
45–41					1
50–46					
55–51				1	
60–56					
65–61				1	
70–66					
75–71				1	

Tables VI to IX (Glenwood Group) show the distribution of differences in months between mental age and achievement ages for the various IQ levels.

Figure 2 shows the same tendency for this group of boys of average range of intelligence, as Fig. I shows for the retarded group. The brightest group have achievement levels much below that which would be expected of children of their mental age. The dullest group, which included children of the same level of intelligence as the brightest level in the retarded group, make scores on the achievement tests much above their mental age. In general, the lower the intelligence, the higher the achievement in relation to the mental age.

IV. DISCUSSION OF RESULTS

From a comparison of Figs. 1 and 2 it is evident that while the children with IQ's from seventy to seventy-nine in the retarded group are making achievement scores from six to thirty months below their mental age in the different subjects, the children of the same level of intelligence in the average group are making achievement scores from two to sixteen months above their mental age. From this it would appear that the relation of achievement to mental age depends not so much upon the level of intelligence but upon the position of that level in the group receiving instruction.

In studies of children in the public schools, Pintner and others[1] have found this same tendency for the school achievement of bright children to be at a lower level than their mental age, and for the achievement of dull children to be at a higher level than their mental age. That is, the accomplishment ratio (EA/MA) of bright children is, on the average, below 1.00 and for dull children it is above 1.00. Pintner[2] gives as an explanation of this that, "In general, children possessing superior intelligence are the ones who are not working up to possible accomplishment, and the final verdict is that our educational system is failing to make use of the vast stores of intelligence which lie hidden and undiscovered." While it is not true that there are "vast stores" of undiscovered intelligence in the retarded group, it is true that the relatively bright children of this group are not working up to possible accomplishment.

The instruction of the teacher is usually adapted to the ability of the middle group or to the ability of the larger number If this is so, the duller children have a greater amount of stimulation relative to their ability than the brighter ones, and the tendency would be for those at the extremes to approach the achievement of the average child This may be one reason why the brighter children are below their mental age in achievement, while the duller ones are above. This, however, should not be so marked in a school where the classes are small.

---

[1] Pintner, R., and H. Marshall: Results of the Combined-mental Educational Survey Test *Journal of Educational Psychology*, Vol. XII, No. 2, pp 82-91

Forgerson, T. L.. The Efficiency Quotient as a Measure of Achievement. *Journal of Educational Research*, Vol. VI, No. 1, pp 22-32.

[2] Pintner, R. "Intelligence Testing," Chapter XI, p 253

TABLE X.—ONE HUNDRED TEN TO ONE HUNDRED NINETEEN IQ

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling
40-44					
35-39					
30-34					
25-29					
20-24					
15-19	2				
10-14	1				
5- 9		2			
0- 4	1	2	1		
5- 1	2	2	1	1	
10- 6	2	2	3	2	1
15-11	3	5	2	2	2
20-16	2	1	1	1	3
25-21		1	1	1	3
30-26	1		3	4	1
35-31	1		1		1
40-36			1	1	1
45-41				1	
50-46					2
55-51					
60-56			1		1
65-61				2	
70-66					
75-71					

Again, the tendency has been to keep children of the same chronological age together in spite of differences in ability. The bright children, therefore, are not required to work up to their possible capacity of achievement, nor are they advanced as far as they should be. The dull children, on the other hand, are urged along in an effort to keep them in their chronological age group, and, because of this insistent pressure, are likely to acquire training above their capacity.

It is also true that the dull children of any given group have had a longer period of training and experience than the bright children of the same mental age in the same group. This training would tend to raise the achievement level of the duller children and retard the achievement of the bright ones in relation to their mental age.

Table XI.—One Hundred Twenty IQ and Above

	Reading paragraphs	Reading words	Arithmetic reasoning	Arithmetic computation	Spelling
40-44					
35-39					
30-34					
25-29					
20-24		1	1	1	
15-19					
10-14	1		1		
5- 9	1	3			1
0- 4	1	2			
5- 1	1	1	1		3
10- 6	2	1	1	1	3
15-11	3	2	1		1
20-16	3	1	1		
25-21	1			2	
30-26	1		4	3	
35-31		1	2	1	
40-36	1	3	2	4	2
45-41			1		1
50-46				1	2
55-51				1	1
60-56		1		1	
65-61	1			1	
70-66					
75-71					1

Tables X to XI (Glenwood Group) show the distribution of differences in months between mental age and achievement ages for the various IQ levels

OBJECTIVE METHODS OF RANKING NURSERY SCHOOL CHILDREN ON CERTAIN ASPECTS OF MUSICAL CAPACITY

THOMAS F VANCE

Iowa State College

AND

MEDORA B GRANDPREY

Ohio State University

The elements in the situation which focused attention upon the possibility of evaluating the musical capacity of nursery school children by a method which might have some approach to reliability were three-fold: an enrollment of thirty-one children, a number of which were almost invariably eager for and interested in the musical phase *of the nursery school program; a teacher whose musical ability* was considerably above the average, with a professional interest in a problem of this kind; and a psychologist on the staff also interested in the problem.

Certain considerations had, of necessity, to be taken into account in outlining the project. The nursery school at the Iowa State College exists primarily as a laboratory for students in Home Economics. Whatever plan of investigation may be adopted must not interfere with the educational program either of the children themselves or of the college students The teacher in this instance was employed to teach the children and not to use them as subjects for research.

Out of these two limitations which the situation inherently forces, an important question arises to which the results of this study may give a partial answer; namely, the potential contribution of the nursery school teaching laboratory to the science of child psychology.

The immaturity of nursery school children is an ever present consideration in projecting a work of this character. In fact, were it not for the child's mental immaturity there would have been no point to the project which is being reported The children would have been rated on their musical capacity according to the Seashore tests. Any method to have any value in determining a child's musical capacity must be exceedingly concrete, must hold his interest and must be within his attention span.

The methods to be described approximate these qualifications They were concrete, they interested the children and they were of short duration. Records were obtained on the following capacities or abilities.

1. Responses to music introduced when the children were engaged in other spontaneous interests.

2. Responses to the music played during the regular music period when the children received some encouragement to take part in it.

3. Imitating the nursery school teacher in singing an interval.

(In the instances just cited, the child was responding as a member of the group. In those that follow, the child was alone with the teacher.)

4. Beating time to graphaphone music with the triangle.

5. General responses to music played on the graphaphone.

6. Imitating the nursery school teacher in beating rhythmical patterns on the triangle.

7. Ratings on the basis of the musical aspect of the home environment.

The observations for responsiveness, Tests Nos 1 and 2, were made on a carefully prepared blank by a student-assistant who knew the children well. They were made two or three days a week for a period of six weeks. A stenographic record was taken for eight different periods. The ratings were based on responses shown through the entire period. A child could be scored as high as five on each of three items, response while playing, presence in circle and response in the circle, with the possibility of a total score of fifteen.

When the graphaphone music was used as a stimulus the teacher simply said to the child, "I have some new records and I thought you would like to hear them." The record was then placed on the machine and the child encouraged to start the machine. The needle was placed in position by the teacher. The teacher then said "Now, see what you can do to that music." Opportunity for three different responses to each selection was allowed before going to the next record but in each case the only encouragement from the teacher was the question: "What can you do to that music?" For no response the score was 0; for verbal response, 1; movements of parts of the body, 2, movement of the whole body, 3; response appropriate to type of selection, 4.

In the phase of the test where the child played the accompaniment to the graphaphone, a plus was recorded for each measure of four beats each of the march and skip rhythms and for each measure of three beats each of the waltz rhythm played in time to the music The skip and march were each sixteen measures in length and the waltz thirty-two measures The records were Victor 20525-B, 20736-A, and 35774-B.

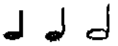
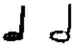
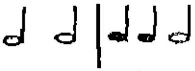
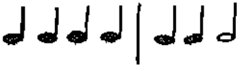
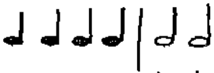
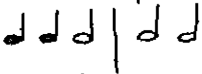
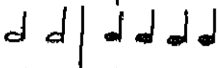
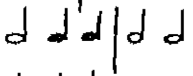
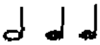
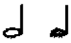
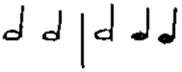
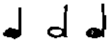
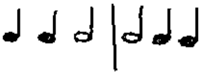
Best results in the singing tests were obtained when it appeared to the children to be a part of the nursery school program At the close of the rest period the name of each child was sung, one by one, with the instruction, "You may roll up your rug." To illustrate, let Jimmie be the first child called. "Jimmie, you may roll up your rug." (Jimmie's name was sung by the teacher) As Jimmie received his direction he would put his rug away and join the teacher at the piano. Another child would then be called, Jimmie following the teacher in singing the name of the second child and so on until all had had their turn. Both ascending and descending intervals were used. Credit was given for responses as follows: Sang one note but off pitch, 1; sang an interval as demonstrated but off pitch, 2; sang an interval of approximate pitch but not on pitch, 3; sang the interval, one note on pitch, 4; sang the interval exactly, 5.

Imitating the nursery school teacher in beating rhythmic patterns on the triangle made the nearest approach to the conventional test situation of any in the series. The patterns themselves, the arrangement, the method of presentation and of scoring which were finally used were all evolved after a rather extended tryout with a three and one-half year old non-nursery school child, a kindergarten child, twenty-one kindergarten children, six kindergarten children and fifty-nine kindergarten children, respectively. There was, however, so little change in the technique of the test after giving it to the fifty-nine kindergarten children that these results may be compared with those obtained from the thirty-one nursery school children who were given the test last of all

The triangle was selected from the other band and orchestral instruments used in the nursery school for several reasons. It is one of the favorite instruments which was played daily in the nursery school by the children themselves. It was given preference to the xylophone because of its constant pitch; to the cymbals, because it could be manipulated more easily; and to the tambourine, because it offered fewer possibilities for variety in playing

The method of presenting the problem to the children was as follows: Immediately before the child was brought in to be tested, the teacher checked her tempo with the beat of the metronome which was turned off before the child came in. When the child came in the teacher began talking about the triangle and showed the child how to hold it. She would grasp it at one of the angles and say, "This is the way I hold it when I play it." Then she would have the child practice taking it, helping him to adjust his fingers to the proper position. In playing she was careful to look at the triangle and to strike it in approximately the same place each time. She then said, "I am going to play some little tunes on the triangle. Then you will play them and make them sound just as I do. Listen." Immediately following, the pattern was played and the triangle and bar placed in the child's hands. "Now, make it sound just as I did," was repeated after the

TABLE I.—ARRANGEMENT OF RHYTHMIC PATTERNS AS GIVEN TO THIRTY-ONE NURSERY SCHOOL CHILDREN WITH TOTAL SCORES FOR EACH PATTERN

C2		49
B3		49
C7		23
C17		13
C11		7
C6		6
C12		5
C5		7
C3		7
B2		3
C8		8
C4		1
C13		0

presentation of the first pattern and as frequently afterward as seemed advisable to keep the attention of the child on the problem in hand.

A response was scored as successful when the child reproduced a pattern once, provided the playing showed a difference of long and short notes in the order presented in the pattern, regardless of accent and volume. If played once in conformity to this standard, the score was counted as 1, if played twice, 2; if three times, 3, and if four times, 4 The child was given ample time to play the pattern four times before the record was made on a prepared blank.

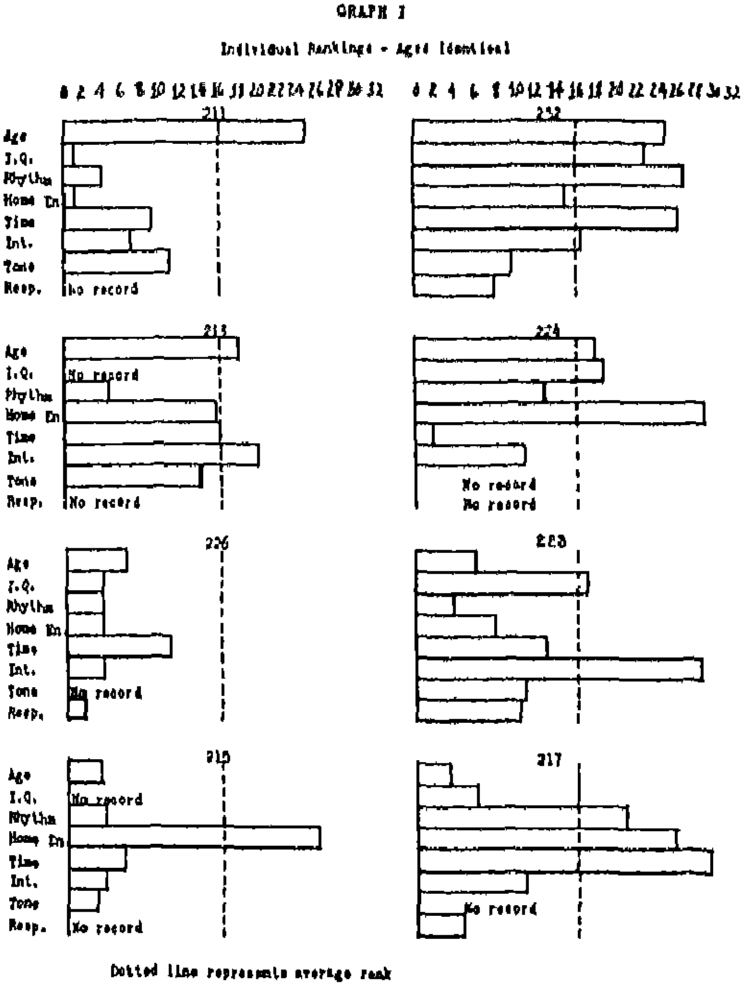
Thirty-six patterns were originally selected from a number of melodies found in a book of songs for young children. In the evolution of the test they were reduced to thirteen as they appear in Table I. It appears from the total scores that a rearrangement is necessary if they are to be given in the order of difficulty.

TABLE II —CORRELATIONS

	IQ	Rhythm	Age	Home Environment	Time	Interpretation	Tone
Rhythm	33 ± 12						
Age		37 ± 11					
Home Environment		28 ± 11	05 ± 12				
Time	− 11 ± 12	30 ± 10	24 ± 11	20 ± 11			
Interpretation	12 ± 12	− 13 ± 12	26 ± 11	12 ± 12	26 ± 11		
Tone	− 04 ± 15	28 ± 13	20 ± 14	10 ± 14	09 ± 15	28 ± 13	
Responsiveness	21 ± 13	41 ± 12	10 ± 11	02 ± 00	20 ± 14	43 ± 12	.52 ± 11

To get a rating on the home environment of the nursery school children as it pertained to music, the teacher went into the homes of the thirty-one nursery school children with a prepared form which she filled out by the method of direct questioning. The form carried questions relative to the provision for musical impression and expression in the home and the musical training and experience in the home and the musical training and experience of the parents. Scores were given on a basis of from one to five credits on each of the following: instruments in the home, singing, training of parents in music and later experience of parents in music.

The limitations of this paper do not permit giving the results in full. The development of methods of determining special aptitudes



GRAPH I.—Individual rankings ages identical

Reading from left to right, individuals of identical ages are grouped in pairs. In the case of the first pair, No. 232 is shown to possess musical capacities of above average or nearly average, while No. 211 falls far below average. No. 224 ranks well above average in time keeping ability, and although he fails to reach the average mark in other capacities, scores well above No. 220 of identical age. A similar comparison may be made in the case of No. 217 and No. 215. Home environment, interpretation and responsiveness to music are abbreviated in the graph respectively as follows Home En., Int., and Resp.

of young children is at present more important than results. The tables and graphs which accompany may be taken as samples

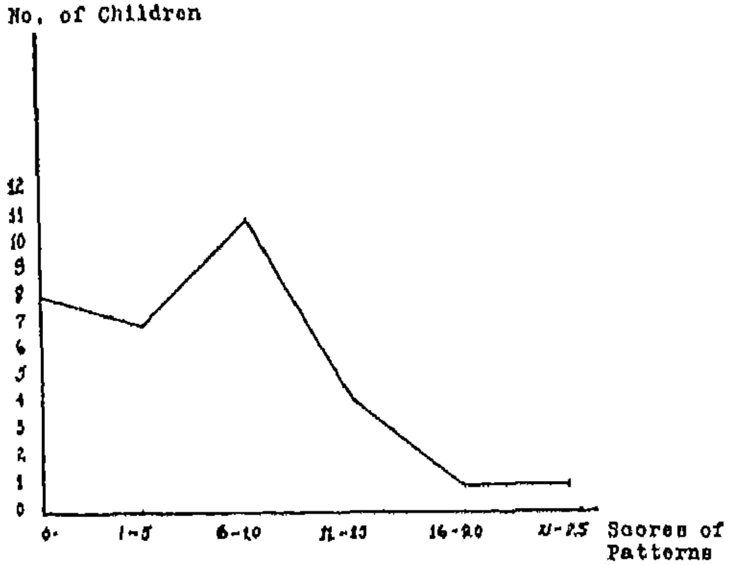
Graph I, giving a comparative picture of eight children paired according to ages, is suggestive of the amount of individual differences in reaction to the various test situations. Whether the assumption that one child has greater musical capacity than another is true can only be determined by a study of the child some ten or fifteen years later. Then there will be the possibility of correlating the results of the work in the nursery school with the results of the Seashore tests With this in view all our nursery school school children are being put through this series, that sufficient data will be available to make correlations significant.

Graphs II and III seem to show that the two tests which they present are better adapted to the older children then to the younger The smaller number in the singing test is due to the fact that the younger children could not be induced to sing They did do something with the rhythm test but eight of them not enough to get a score of more than zero The tests are not recommended for children younger than three years In the rhythm test the median for the kindergarten children was 14 8, for the four to five year level, 8.84; the three to four year level, 4.18 and for the two to three year level, 2 43 Further, the correlation of rhythm with age in the kindergarten was .03 while in the nursery school it went as high as 37 To give any test of special aptitude too early may give merely a measure of maturity and not of the aptitude in question.

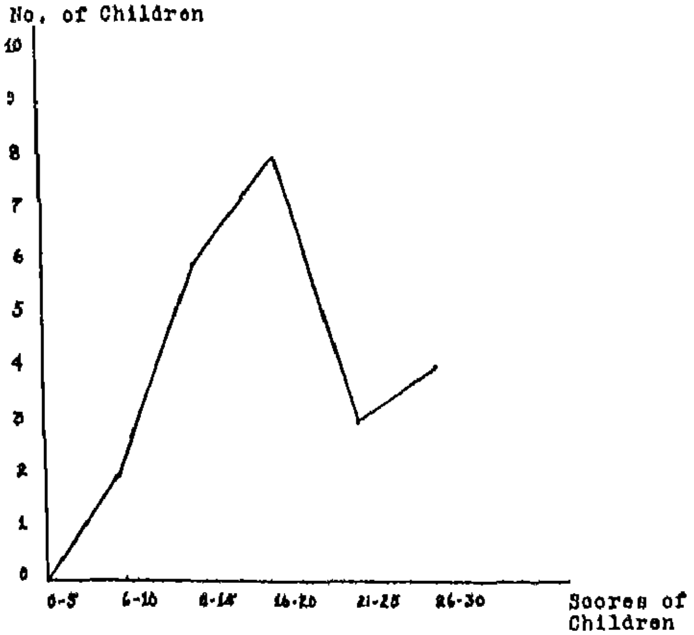
The highest correlation to be found in the table is .62 between responsiveness and home environment. It may suggest a possible carry over of musical expression in the home to expression in the nursery school, making perhaps some slight allowance for inheritance

One or two by-products seem to have something more than passing interest. In the singing test the average scores for the descending intervals are higher than those for the ascending intervals. The descending intervals were also lower in pitch. The question arises, "Would this difference still obtain if pitch were held constant and a much larger number of cases studied?"

In the rhythm test the patterns in which the quarter notes preceded the half notes received higher scores than those of the reverse order This difference obtained in both the kindergarten and the nursery school groups.



GRAPH II.—Distribution of scores made by 31 nursery school children on rhythmic pattern test.



GRAPH III.—Distribution of scores made by 23 nursery school children in interval singing test.

Taken as a whole, the study may be used as an illustration of the use of the nursery school as a teaching laboratory, as differentiated from a research laboratory, for securing data which may prove of scientific as well as of practical value. Practical results certainly accrue in the stimulation of the interest of the nursery school staff to secure as reliable information as possible about the children. Only the future can determine the scientific value of such a study.

FACTUAL MEMORY OF SECONDARY SCHOOL PUPILS FOR A SHORT ARTICLE WHICH THEY READ A SINGLE TIME[1]

ALFRED G. DIETZE AND GEORGE ELLIS JONES

University of Pittsburgh

INTRODUCTION

The Problem. A review of the literature reveals that but few of the many studies of memory touch upon the subject from the point of view of what is actually remembered of things experienced in ordinary life situations. The writers feel that research in this general field should give profitable returns and have undertaken an investigation of a typical memory situation, namely, that of a single reading of a short informative article by pupils in the secondary schools. The present study is an attempt to determine the range and average extent of immediate and delayed factual memory of pupils in Grades VII to XII for a short article which they read a single time and the relation which obtains between memory at intervals of increasing length.

This and related problems emerge in a consideration of the practical significance of the classical laboratory studies of memory. It is uncertain, for example, whether we remember what we obtain by reading and study in the same way that we remember material learned by rote to the degree of one errorless and unhesitating recital. The pedagogical significance of exact information upon this point relative to a single reading of an article is clear in view of the fact that pupils in school and college commonly prepare assignments by reading lessons a single time.[2]

Related Studies.—A number of related investigations have been reported during the past ten years. Jones studied the memory of college students for the content of lectures in psychology.[3] The

---

[1] Acknowledgment is made to Professor H. B. Reed for fundamental suggestions.

[2] *Cf.* Yoakam, G. A.: "Reading and Study." The Macmillan Co., New York, 1928, pp. 185–190.

[3] Jones, H. E.: Experimental Studies of College Teaching. *Archives of Psychology*, No. 68 (1923).

students knew an average of 62 per cent of the facts presented when they were tested immediately after the lectures. After three or four days they remembered an average of 45 per cent, and after eight weeks memory dropped to 24 per cent.

Several writers report conflicting results regarding memory over the summer vacation.[1] Recently Bassett published a study of the retention of history in the sixth, seventh and eighth grades.[2] Objective tests were given at the end of the course and again after intervals of four, eight, twelve and sixteen months. Computing retention on the basis of the scores in the initial tests, average memory scores of 86.01, 81.51, 76.72 and 71.96 per cent respectively were obtained. Since, however, no measures of validity and reliability are presented, and since the tests are short and include four different types of questions to which equal credit was given in the scores, the validity of the conclusions may be questioned. At least the method hardly justifies a statement such as, "It is worthy of note that after sixteen months the children know seventy-two per cent of the history which they knew at the end of the semester."[3]

In respect to method and results there are many points of similarity between Yoakam's study of the effects of a single reading and the present investigation. His finding that the effect of a single reading gives on the average less than half of the total ideas in the article[4] is of especial interest here. He also found that that the effect of a single reading varies with the individual and from grade to grade; and he questions whether a single reading without any immediate recall will leave any impression on the mind of the learner after a lapse of twenty or thirty days, unless the material is very interesting or striking.

---

[1] Garfinkle, M. A. The Effect of the Summer Vacation on Ability in the Fundamentals of Arithmetic. *Journal of Educational Psychology*, Vol. X, 1919, pp. 44–47; Kirby, T. J. "Practice in the Case of School Children." Columbia University Contributions to Education, No. 58. Bureau of Publications, Teachers College, Columbia University, New York, 1913; Noonan, M. E. "Influence of the Summer Vacation on the Abilities of Fifth and Sixth Grade Children." Columbia University Contributions to Education, No. 204. Bureau of Publications, Teachers College, Columbia University, New York, 1926.

[2] Bassett, S. J. "Retention of History in the Sixth, Seventh and Eighth Grades with Special Reference to the Factors that Influence Retention." The Johns Hopkins University Studies in Education, No. 12. Johns Hopkins Press, *Baltimore*, 1928.

[3] See *ibid.*, p. 51.

[4] Yoakam, G. A.: The Effects of a Single Reading. *University of Iowa Studies in Education*, Vol. II, No. 7. The University of Iowa, Iowa City, 1924. See especially conclusions on page 98ff.

MATERIALS AND METHOD

Subjects.—Children in Grades VII to XII acted as subjects, a total of 2789 taking part. The experimental groups consisted of 2062 pupils in the junior and senior high schools of Uniontown, Pa. Each grade was divided into six numerically approximately equal groups for purposes of rotation in the experiments. The remaining seven hundred twenty-seven subject, drawn from neighboring schools, were used in control experiments: Control Group A, two hundred seventy-eight children, for obtaining intercorrelations between the tests; Control Group B, two hundred forty-five pupils, for an investigation of previous knowledge of the articles; Control Group C, one hundred eighty-two pupils, for obtaining a measure of how well pupils comprehend the materials. The median scores of these groups in intelligence tests, reading and vocabulary tests correspond closely to published norms; hence, it may be safely assumed that they are fairly typical of the secondary school population.[1]

The Reading Selections.—Three interesting and highly factual articles were used as reading selections, each printed in the form of a four-page pamphlet. "Radium: The Magic Metal," containing 1265 words, is an account of the discovery and uses of radium. "The Early Germans" describes the habits and customs of the early inhabitants of central Europe and is 1061 words in length. "Sir Richard Arkwright" is an account of the life of the English inventor and contains 1279 words. These articles are well suited to the purpose for which they were selected, since they abound in facts upon which questions can be formulated and are of suitable difficulty for the grades studied.[2]

---

[1] *Complete tables for all summarizations contained in this paper are to be* found in A. G. Dietze, "Factual Memory of Secondary School Pupils for a Short Article Which They Read a Single Time." Doctor's dissertation, Library of the University of Pittsburgh, 1930.

[2] For methods of determining difficulty *cf. ibid.*, pp. 51–56

The Tests—The tests were made up of five-response multiple choice type questions and were printed in the form of four-page pamphlets. An effort was made to thoroughly canvass the articles for all thought units contained so as to make the memory tests as complete as possible. The test on Radium contains one hundred items, the other two one hundred twenty-five items each.

In respect to validity, objectivity and reliability these tests compare favorably with some of the better tests in standard use. Validity is attained by thoroughly covering the content of the articles in the questions and is rendered evident by an average intercorrelation, corrected for attenuation, of .80 ± .014 between the tests. Objectivity is attained by (*a*) careful instruction for administration, (*b*) clear directions to pupils, (*c*) unambiguous questions, (*d*) uniform and well marked scoring keys in the form of strip stencils, and (*e*) supervision of scoring by one responsible person. The reliability of the tests is shown by the following figures, the reliability coefficients being obtained by applying the Spearman-Brown prophecy formula to the correlation between odd and even halves of the tests:

	Radium	Germans	Arkwright
Reliability of half test	82 ± 008	80 ± 005	85 ± 008
Reliability coefficient	00 ± 005	04 ± 003	02 ± 004
Index of reliability	95 ± 002	97 ± 001	96 ± 002
PE of estimate	3 63	3 00	3 72
PE of measurement	2 63	2 50	2 68
Number of cases	650	607	616

Scoring—Since the tests are designed to measure the amount retained, the best type of score seems to be the percentage of items answered correctly. In this way each item is considered a unit, relative difficulty of the separate questions being disregarded. No attempt is made to correct for guessing, since instructions were given not to guess. It may be stated that this method of scoring resulted in distributions approximating the normal about as well as can be expected from the number used. Arkwright gave an especially good fit to the normal curve.

Procedure.—The method of conducting the experiments was to give the pupils an article, with instructions to read it a single time in order to answer a memory test at some time later. The pupil was allowed to read at his own rate. The test was given immediately after the reading in Experiment 1, and after intervals of 1, 14, 30 and 100 days in Experiments 2, 3, 4 and 5 respectively. Each group participated in the immediate memory experiment and in two delayed memory experiments, using different materials for each. The manner in which the groups and materials were rotated appears in the following scheme:

Group	Experiments participated in and material used in each		
I	Exp. 1, Radium	Exp 2, Germans	Exp 5, Arkwright
II	Exp. 1, Radium	Exp. 3, Germans	Exp 4, Arkwright
III	Exp. 1, Germans	Exp. 2, Arkwright	Exp 5, Radium
IV	Exp. 1, Germans	Exp. 3, Arkwright	Exp. 4, Radium
V	Exp. 1, Arkwright	Exp 2, Radium	Exp 5, Germans
VI	Exp. 1, Arkwright	Exp 3, Radium	Exp. 4, Germans

The reading and testing periods were administered by the English teachers of the regular classes, thus providing a natural school-room situation to the subjects. The teachers were carefully coached in the manner of conducting the experiments and instructed to report all irregularities to the experimenter. Before the experiments proper were begun, all groups were given a practice exercise consisting of a reading selection and test similar in nature to those above described. This served to acquaint the subjects thoroughly with the details of taking the tests.

SUPPLEMENTARY RESULTS

Several factors which cannot be eliminated by our method influence the results obtained. The approximate extent to which two of these enter was made the object of two control experiments. The points investigated are (*a*) the influence of previous knowledge and ability to answer questions by logical thinking and (*b*) the failure of pupils to answer questions because of comprehension difficulties and careless work habits.

Previous Knowledge.—Control Group B was required to answer the tests without previously having read the articles. They were, furthermore, instructed to guess when they did not know the answer to a question. The results are somewhat confusing, since the medians are in inverse relation to grade. The reason for this, however, becomes apparent upon examination of the papers. The lower grades complied more fully with the directions regarding guessing and, therefore, had more opportunities of getting items correct by chance. The actual medians range from 32.3 per cent in the seventh grade to 23.5 in the twelfth for Radium, from 30.8 in the seventh to 25.0 in the eleventh for Germans, and from 24.8 in the seventh to 18.5 in the twelfth for Arkwright.

The instructions regarding guessing were given for the purpose of extracting the correctly guessed items from the scores by the commonly used formula for correcting multiple choice tests;[1] however the correction could be accomplished only in the seventh grade, since in this grade the directions were strictly followed. When the scores for the seventh grade are subjected to the formula, the following results remain as the approximate scores that pupils in this grade are able to make without previously reading the articles: Radium 15.4 per cent; Germans 13.4 per cent; Arkwright 6.0 per cent. These amounts are interpreted as due to previous knowledge and logical thinking and are in harmony with the findings of Yoakam in a similar experiment.[2]

Comprehension Difficulty and Carelessness.—The ability of pupils to answer the tests when they are allowed to look up the answers in the articles may be taken as a measure of how well they comprehend the materials and how careful they are in their work. Control Group C was used for determining this factor. The means for the respective grades range from 82.0 per cent in the eleventh grade to 66.0 in the seventh for Germans, and from 73.0 in the twelfth to 54.3 in the seventh for Arkwright. The difference between grades are for the most part reliable. The average of all grades in Germans is 75.7, in Arkwright 63.5. Thus this grade missed about twenty-five per cent of the questions on Germans and about thirty-seven per cent of those on Arkwright because they did not comprehend the materials read or were too careless to do thoroughly accurate work

---

[1] Corrected score $= \text{No. right} - \dfrac{\text{No. wrong}}{(n-1)}$

[2] *Op. cit.*, pp. 58–59

Although the influence of the above factors have not been determined with exactness, nevertheless the control experiments show that it is significant; and the experimental results must, accordingly, be interpreted by the reader with these facts in mind.

FACTUAL MEMORY AFTER FIVE INTERVALS

Statistical Treatment. -Two methods of summarizing the results are employed. The first makes use of the common statistical device of describing group results in terms of means and standard deviations. This method brings out the trend of the data in a straightforward manner which needs no further comment.

The other method is made necessary because the technique of equated parallel groups was not followed. For this reason the various sub-groups are only roughly comparable, the group average depending to a large extent upon the ability of the particular group relative to entire grades or all groups taken together. To obtain comparable results for grades or other sub-groups, it is necessary first to transmute scores derived from the three different materials into equivalent terms and then to combine the results into an average representative of the combined groups.

The best method for accomplishing this is to make use of the regression equations derived from the correlations between the materials. It will be remembered that Control Group A was used for the purpose of getting correlations between the three materials. From these correlations the regression equation between Radium and Arkwright and between Germans and Arkwright were derived, the resulting equations making possible the conversion of results in Radium and Germans into scores equivalent to Arkwright scores and the combination of all three groups in terms of an average for the latter material. To distinguish this derived score from the scores of groups actually using Arkwright the term "A-score" will be used. The actual equations for deriving A-scores from the results in Radium and German were found to be:

$$\text{A-score} = .80 \text{ Radium} + 2.4 = .67 \text{ Germans} + 11.0.$$

Similar scores may, of course, be derived in terms equivalent to Radium or Germans. The reader may get an approximate value for R-scores or G-scores by adding ten points to the A-score averages herein reported.

Immediate Memory—Table I gives the means and standard deviations of the respective groups for the material which they read and the A-score averages of the combined groups in immediate memory

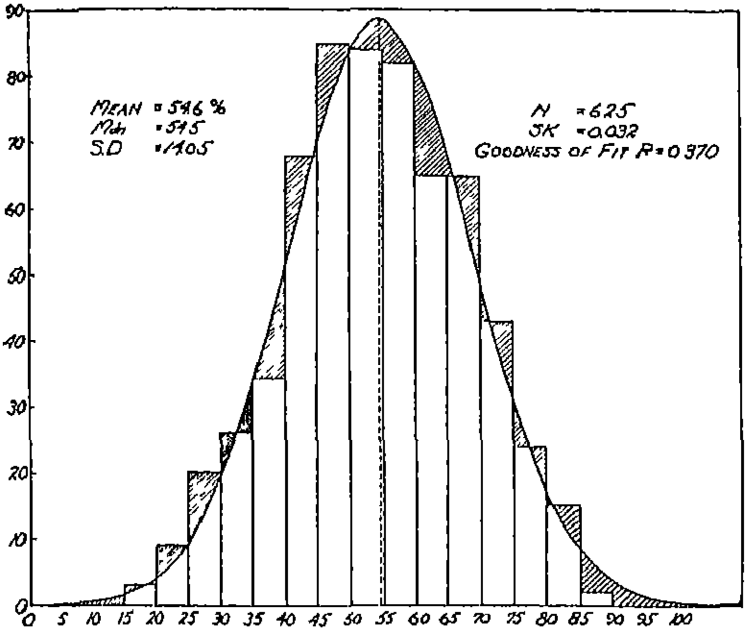


FIG 1 —Histogram for immediate memory of Arkwright compared with a normal curve fitted to the data

The total range of the distributions are from thirty-one to one hundred per cent for Radium, from sixteen to one hundred per cent for Germans, and from sixteen to ninety per cent for Arkwright. Thus pupils in the junior and senior high school score over four-fifths of the entire scoring range of these tests, which shows that individual differences in this ability are great. The distributions for Radium and Germans are somewhat negatively skewed, *Sk* being −.352 and −.265 respectively [1]

---

[1] $Sk = \dfrac{3(\text{Mean} - \text{Median})}{\text{SD}}$

This indicates that there is some piling up of immediate memory scores in the upper part of the range for these articles. Arkwright, however, gives a very good approximation to the Gaussian curve, Sk being only .032 and Pearson's chi-square test of goodness of fit giving a P value of .370, which is very satisfactory for a population of slightly more than 600.[1] The comparison of the latter distribution with the theoretical normal curve fitted to the data is shown graphically in Fig. 1.

TABLE I.—AVERAGE PERCENTAGE IMMEDIATE MEMORY

Grade	Radium			Germans			Arkwright			A-Score	
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	N
XII	84.3	7.0	68	75.8	12.5	67	61.8	11.2	100	65.2	244
XI	77.0	10.1	116	73.0	12.2	101	62.7	11.2	61	62.3	278
X	76.0	11.6	121	66.0	12.3	122	57.4	11.9	123	59.0	366
IX	72.4	10.6	139	67.3	13.6	109	59.3	12.1	126	55.5	374
VIII	66.6	12.1	72	57.9	13.5	120	50.0	14.1	131	51.3	323
VII	67.3	12.6	146	53.8	15.0	138	43.0	12.0	75	50.0	359
VII–XII	73.4	12.4	662	64.8	15.7	657	54.6	14.1	625	56.7	1944

TABLE II.—PERCENTAGE MEMORY AFTER ONE DAY

Grade	Radiums			Germans			Arkwright			A-Score	
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	N
XII	72.9	10.7	74	71.8	9.6	30	63.3	8.9	35	61.1	139
XI	64.9	11.8	26	59.3	14.9	53	55.2	9.4	52	53.2	131
X	56.3	10.7	43	58.8	15.7	60	46.7	12.7	57	47.9	160
IX	56.1	13.9	62	54.2	14.8	68	42.4	12.8	31	46.0	159
VIII	55.0	12.3	61	55.1	13.0	34	43.4	10.3	57	45.3	152
VII	47.7	16.4	38	46.0	16.3	69	33.5	8.3	71	40.0	178
VII–XII	58.7	16.1	301	56.7	17.2	312	46.4	15.6	303	48.0	919

Delayed Memory, One Day.—The average results for memory after one day are given in Table II. The distributions are similar in nature as in the previous experiment. The range in Radium is from 11 to 95 per cent, in Germans from 16 to 95, and in Arkwright from 6 to 85.

---

[1] See K. J. Holzinger, "Statistical Methods for Students of Education" Ginn and Co., Boston, 1928, pp. 245–248.

Thus the same wide range of individual differences is to be noted as in immediate memory.

Delayed Memory, Fourteen Days.—Results for memory after fourteen days are shown in Table III. The distributions at this interval have become positively skewed, showing that the task is becoming more difficult. *Sk* is .191 for Radium, .405 for Germans and .405 for Arkwright. The range in Radium is from eleven to seventy per cent, that in Germans is from six to eighty-five per cent, and that in Arkwright from one to seventy per cent. Again we find ability in these grades to range from very low to very high.

TABLE III.—PERCENTAGE MEMORY AFTER FOURTEEN DAYS

Grade	Radium			Germans			Arkwright			A-Score	
	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	*N*
XII	40.1	11.8	34	51.3	13.2	27	32.0	10.0	27	30.8	88
XI	50.0	11.4	31	46.7	13.8	54	32.3	12.3	42	30.0	127
X	40.8	10.9	74	44.2	17.6	48	34.4	12.1	56	30.1	178
IX	44.9	9.7	64	38.8	11.0	62	32.9	10.9	69	35.9	195
VIII	45.2	10.6	64	34.3	9.0	27	28.8	11.3	56	33.8	147
VII	40.7	9.5	35	37.9	10.9	71	37.5	15.2	59	36.3	165
VII–XII	43.7	11.0	302	42.0	14.1	289	33.1	12.0	309	36.3	900

TABLE IV.—PERCENTAGE MEMORY AFTER THIRTY DAYS

Grade	Radium			Germans			Arkwright			A-Score	
	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	*N*
XII	37.9	9.8	23	30.3	9.3	35	31.3	11.0	30	33.8	88
XI	42.1	12.6	50	40.1	11.0	26	31.5	12.5	48	34.0	124
X	42.2	11.1	54	32.1	8.5	34	37.8	14.4	47	35.7	135
IX	37.1	11.5	65	32.1	9.1	53	25.4	8.9	67	29.5	185
VIII	37.7	11.1	58	41.3	11.4	61	27.6	7.7	26	33.2	145
VII	37.2	11.1	56	32.0	9.5	34	25.9	9.0	68	25.5	168
VII–XII	38.3	11.0	306	36.4	10.7	243	30.3	12.0	286	32.4	835

Delayed Memory, Thirty Days.—Average results for memory after thirty days are shown in Table IV. The nature of the distributions (skewness) is the same as in the preceeding experiment Scores in Radium range from six to eighty per cent, in Germans from six to seventy-five, and in Arkwright from one to seventy-five.

Delayed Memory, One Hundred Days The results for the fifth experiment are shown in Table V. The scores in Radium range from one to seventy-five per cent, in Germans from one to sixty, and in Arkwright from one to ninety As in the previous experiment, skewness is slightly positive, showing massing of scores below the mean.

TABLE V - PERCENTAGE MEMORY AFTER ONE HUNDRED DAYS

Grade	Radium			Germans			Arkwright			A-Score	
	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*N*	Mean	*N*
XII	39.1	9.8	32	33.3	11.7	64	22.6	16.9	25	31.2	121
XI	35.1	11.9	47	35.0	7.6	29	23.0	18.1	26	29.6	102
X	30.6	6.8	45	31.4	10.2	43	28.7	9.6	59	30.5	147
IX	38.0	8.2	30	26.6	8.7	27	21.9	10.3	33	28.5	90
VIII	31.0	8.3	57	29.4	7.5	55	22.0	15.4	29	27.9	141
VII	27.3	7.1	68				20.5	7.0	63	25.6	131
VII–XII	33.0	9.8	279	32.1	9.6	218	23.4	12.2	245	30.0	732

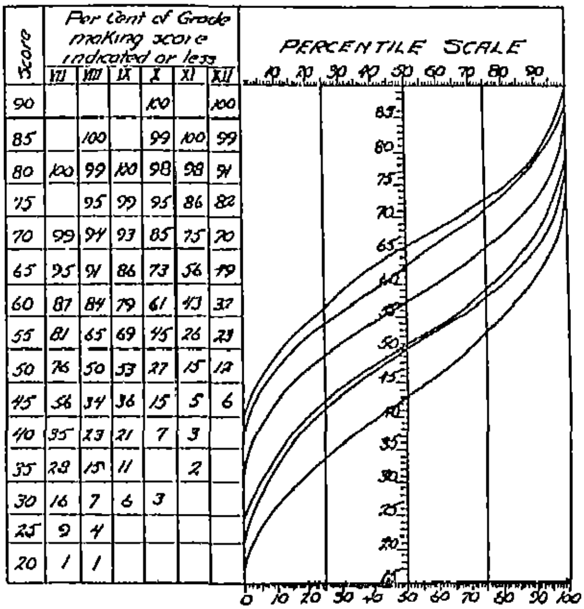
COMPARISON OF GRADES IN FACTUAL MEMORY

Grade Averages. The data presented in the previous Tables (I–V) enable us to make the following comparisons of the grades participating in these experiments.

In immediate memory the averages increase quite markedly from grade to grade. Of the fifteen comparisons that may be made on the basis of the results of the various materials eight of the differences turn out to be four times their probable errors or more and are, therefore, completely reliable. Three more comparisons are differences 1.8 times their probable errors, or better, the chances of a true difference greater than zero being at least ninety in one hundred A comparison of the A-Score means reveals an average increase of 3 05 per cent from grade to grade, all the differences being reliable

In memory delayed one day the grade averages again increase as in Experiment 1. Of the fifteen comparisons six give reliable differences, and four more are more than 2.5 times their probable errors The A-Score differences between the means of the grades average 4.5 per cent, but all are not completely reliable.

In the experiments with intervals of fourteen days and longer, no relationship is to be noted between factual memory and grade, the rank order of the grades being haphazard and most of the differences being negligible. Whereas the means of the seventh and twelfth grades in the immediate memory and one day delayed memory experiments are well set apart by a distance of about 1 5 standard deviation of the entire distribution, the differences between the lowest and highest averages in the experiments with longer intervals are small It may be inferred that groups originally quite different with respect to knowledge of an article read a single time become more and more alike with lapse of time—*i.e.*, the factors influencing forgetting operate differently on groups of differing abilities.



Score	Per Cent of Grade making score indicated or less					
	VII	VIII	IX	X	XI	XII
90				100		100
85		100		99	100	99
80	100	99	100	98	98	91
75		95	99	95	86	82
70	99	94	93	85	75	70
65	95	91	86	73	56	49
60	87	84	79	61	43	37
55	81	65	69	45	26	23
50	76	50	53	27	15	12
45	56	34	36	15	5	6
40	35	23	21	7	3	
35	28	15	11		2	
30	16	7	6	3		
25	9	4				
20	1	1				

FIG. 2—Percentile curves showing the overlapping of grades in immediate memory for Arkwright

Overlapping of Grades—While the average score in immediate and one day delayed recall increases perceptibly with grade, nevertheless adjacent grades overlap to a considerable extent We find, *e.g.*, that the seventh, eighth, ninth, tenth, eleventh and twelfth grades reach or exceed the mean in immediate memory of Radium for all grades taken together by the following amounts respectively: 33 6 per cent, 25.0 per cent, 40 3 per cent, 57 0 per cent, 58.6 per cent, and 86.8 per cent. Again, 85 6 per cent of the distribution of the combined grades for Radium in immediate memory fall within a range common *to all six grades*, and, similarly, 87.2 per cent of the combined scores in Germans and 82.6 per cent of those in Arkwright. The degree of overlapping is strikingly brought out by the percentile graph for immediate memory of Arkwright, Fig. 2, from which may be read the percentage of scores of any grade which fall below or above any given point in the distribution of any other grade. The curves have been smoothed free-hand.

Correlations.—The product moment correlations obtained for the factual memory results of several of the groups with grade in school are presented in Table VI. It will be seen that the correlations between immediate memory and grade are positive and low, those between delayed memory and grade are negligible. These results are in accord with the foregoing analysis of grade averages. The slight correlations which do exist, as will be shown in a later article, are due to other factors rather than the accident of grade.

TABLE VI.—CORRELATIONS OF FACTUAL MEMORY AND GRADE IN SCHOOL

Variables	Radium		Germans		Arkwright		Average
	r	PE	*r*	PE	*r*	PE	r
Immediate memory and grade	40	035	24	011	29	045	34
Fourteen day memory and grade	04	010	34	010	− 07	013	10
Thirty day memory and grade	11	.013	07	019	30	000	18

(*To be concluded in December issue*)

WHY OTIS' "IQ" CANNOT BE EQUIVALENT TO THE STANFORD-BINET IQ

PSYCHE CATTELL

Harvard University

Dr. Otis has devised a measure of brightness for use with his Self-administering Tests which is derived by determining the number of points in score the child varies from his age norm, if he is above the norm this amount is added to one hundred, if below subtracted from one hundred. Dr. Otis states that when an IQ has been found by other means than the Binet Tests the term "IQ" must not be used without qualification, yet he writes that "it seemed best simply to call this new measure of brightness an 'IQ.'" He apparently justifies the use of the term "IQ" on the erroneous belief that his measure of brightness is equivalent to the Binet IQ and that the group test IQ's obtained by the standard method are not.

Dr. Terman found the middle fifty per cent of the Stanford-Binet IQ's *to have a range from ninety-two to one hundred eight*. Dr. Otis found the middle fifty per cent of the scores made in his tests to have approximately the same range in the several age groups, that is that fifty per cent of the cases fell within eight points of the norm, he then concluded:

> Fortunately, therefore, each point in the score of an individual above or below the norm for his age represents a point in IQ above or below one hundred. If an individual score exceeds the norm for his age by twelve points, his IQ is one hundred twelve . . . . This being the case it seemed best simply to call this new measure of brightness an "IQ."

Dr. Otis has failed to take into account the fact that two distributions may have the same medians and the same middle fifty per cent range and still differ widely at the extremes. It can be shown both theoretically and empirically that there is a constant difference between the Binet IQ and Otis' "IQ" at the extremes, especially the upper extreme. That this must of necessity be true can be demonstrated from the data given in the manual of directions for the "Otis Self-administering Tests of Mental Ability." An inspection of the table of norms to determine the highest and lowest "IQ" that it is possible for a child of a given age to obtain by this method will show the impossibility of the range of Dr. Otis' measure being equal to that of the Binet IQ.

While Dr. Otis makes no mention of the relative reliability of his table of norms at different ages, it is perhaps, not fair to expect too great accuracy at the extremes. Therefore, the ages of 10-6, 13-0 and 15-6 have been selected for a comparison of the Binet IQ and the "IQ" obtained by Dr. Otis' method. These ages are the quartile points of the age range of the table of norms and should include the more reliable part. The curve to the left of the accompanying chart is the frequency distribution of the IQ's of Professor Terman's original one thousand unselected Stanford-Binet IQ's. The curve to the right is a frequency distribution of the Binet IQ's of his six hundred forty-three gifted children, largely from Grades III to VIII and all below the age of fourteen years The lowest and highest Otis "IQ's" that it is possible for children of a given age to obtain on the Intermediate Self-administering Test are indicated on the IQ scale at the foot of the chart.

The highest possible Otis "IQ" that could be obtained by a child aged 13-0 on this test is one hundred thirty-one and the lowest fifty-six. This is a narrower range than that of the Stanford-Binet IQ's. If Professor Terman had used this test and method in selecting his gifted group he might have found one or two children that measured up to one hundred thirty but none above one hundred thirty-one (his group of six hundred forty-three contained only one case with an IQ below one hundred thirty-five). At the age of 10-6 the highest possible Otis "IQ" is one hundred forty-eight (approximately one half of Professor Terman's gifted group had IQ's of one hundred fifty or higher). The lowest a child could attain at this age is seventy-three Every large school system has a number of children with Binet IQ's below seventy, At age 15-6 the highest "IQ" obtainable is only one hundred twenty approximately two standard deviations above the norm. Even if the advanced test were substituted for the intermediate the highest possible "IQ" would still be only one hundred thirty-seven. The lowest "IQ" obtainable at this age is forty-five, even this is not as low as the Standford-Binet IQ's frequently met with. At no age is there a possible Otis "IQ" range of more than seventy-five points. At the age of 12-0 it is theoretically possible to get a range from sixty-three to one hundred thirty-eight. A younger child could get a higher Otis "IQ," but could not make one as low as sixty-three, while an older child could make one lower than sixty-three,

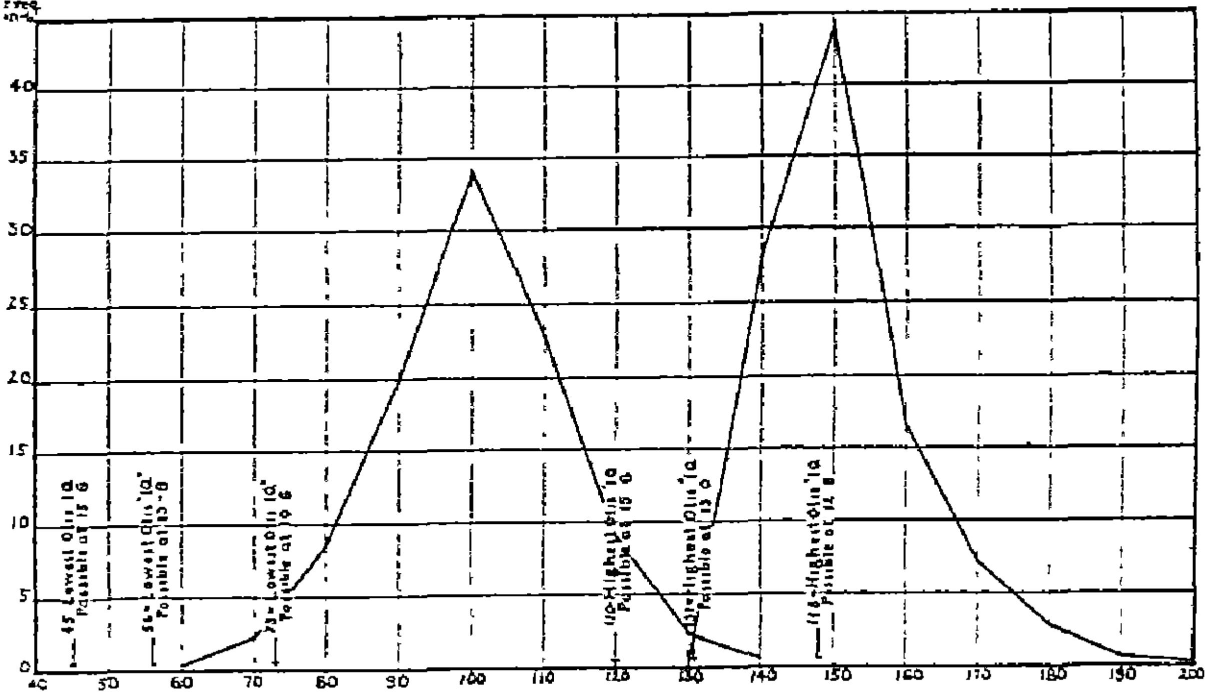


FIG 1—Why Dr. Otis "IQ" cannot be equivalent to the Stanford-Binet IQ.

but would have no possibility of making one as high as one hundred thirty-eight.

Similar conclusions were reached in an earlier study[1] based on the actual IQ's obtained by giving Forms A and B of the Otis Self-administering tests and at least two Standard-Binet tests to three hundred twenty-two school children ranging in age from ten to thirteen inclusive. Those children who according to the Stanford-Binet tests possessed an average grade of intelligence were found to be about six points lower on the average in Dr. Otis measure in Form A and two points above in Form B. At the lower levels of intelligence the differences are slightly less, but at the higher levels as selected by the average of two or more Stanford-Binet IQ's the Otis measure is markedly and consistently below that of the Stanford-Binet, especially in Form A where the average difference was about twelve points at the Binet IQ level of one hundred twenty to one hundred thirty and about seventeen points for that above one hundred thirty. *These are median differences, there were, of course, many* larger individual differences, fifteen per cent of those with Binet IQ's above one hundred twenty-four varied by twenty-five or more points, all in the negative directions.

Thus the Otis "IQ" is seen to differ from the Binet IQ by an *appreciable amount near the norm and to differ markedly at the* upper extreme. It is difficult to see any justification for applying the unqualified term "IQ" to a measure which differs from the Standard IQ to such an extent. It is strange that Dr. Otis should propose doing so when he writes on the same page that:

> *It seems that the term "Intelligence Quotient" is coming to have a legal recog*nition, but IQ's as sometimes derived from group tests of mental ability bear little relation to IQ's derived by the Binet Tests[3] . Unless it is distinctly understood how IQ's were derived . they should be designated by some means such as National IQ's, Otis IQ's or Binet IQ's. The term "IQ" when not so qualified or understood, must be interpreted as referring to actual Intelligence Quotients found by means of the Binet Tests[3] . It is the purpose of the author to use the term "IQ" only in its original significance[2]

---

[1] CATTELL, PSYCHE. IQ's and the Otis Measure of Brightness. *Journal of Educational Research*, June, 1930.

[2] "Otis Self-administering Tests of Mental Ability." Revised Manual, p 5.

[3] "Statistical Methods in Educational Measurements." World Book Co., Yonkers, New York, p. 156.

Yet in proposing his new measure of brightness which has just been shown to differ from the IQ both in its meaning and method of derivation he apparently believes that he is adhering to his statement that "It is the purpose of the author to use the term 'IQ' only in its original significance" when he writes: ". . . It seemed best simply to call this new measure of brightness an 'IQ.'"

PROGNOSIS OF ABILITIES TO SOLVE EXERCISES IN GEOMETRY

WINONA M. PERRY

University of Nebraska

What are the abilities essential to the solving of the various exercises confronting the student of plane geometry? Since it is possible that these abilities may be component parts of "general intelligence," analysis of intelligence tests might be the means of detecting the constituent abilities. Accordingly, twenty-six group intelligence tests (or forms) were examined and their content observed as verbal, numerical, or spatial in character. The separate elements, and subtests too, were analyzed further into (1) abilities to perceive singly (as in recognition of the name, size or location of an object), or to perceive several objects for the purpose of comparisons of similar or non-similar meanings, shapes, or amounts (as in scaling to denote relative amounts of beauty); into (2) abilities to recall word meanings, attributes, and relationships; and into (3) abilities in purposive thinking: To discover relationship (as in number series, analogies, and mixed relations, disarranged elements, completion of sentences and pictures), to judge the possibility and correctness of statements, to apply a principle singly, and to weight elements (as in finding the best answer, the correct pathway, and in dealing with certain elements inferentially).

Which of these abilities are selected and organized into the series of changing responses occurring during the processes of solving geometric exercises? Perception is essential; perception, in terms of the given figures as a whole or in parts, is dependent upon the recall of meanings of figures, of relationships, and of properties, *i.e.*, a sufficient degree of meaning or familiarity with vocabulary units is necessary to note and to follow commands or questions in terms of the total figure as given, or its parts, or the relationship between these parts. Not essential, but usually following this series of perceptions is their motor expression in the drawing of the figure described and in the stating of the facts given and those which are to be proved; this combination of responses results in perception of known relationships in the figure as drawn. Recall in the light of analysis means pertinent recall which immediately makes available the needed meaning of certain figures previously reacted to, or their nature, relationships, or properties. Very complicated is the series of responses, "purposive thinking," which is to determine the link that connects the conclusion—to be judged, drawn, or proved—with the recalled and analyzed facts. The more major responses included under purposive thinking may be indicated by these questions:

Which elements among those recalled "fit in" with the conclusion?

Which elements among those analyzed in the light of the figure "fit in" with the conclusion?

Which ones among the pertinent recalled elements differ from, or are not included among the pertinent analyzed facts?

Is evidence available to help in finding *the one* element necessary to complete the identity between the series of analyzed and of recalled facts?

S - - - - - - → R_1

[If a diagonal of a quadrilateral bisects those angles whose vertices it joins, the diagonal divides the figure into two equal triangles]

[Meaning of statement as a whole]

R_{11} [Perception of conclusion, that triangles are equal]

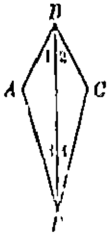
R_{12} [Perception of total figure, a quadrilateral.]

R_{13} [Perception of relationship between parts of figure, diagonal bisecting opposite angles]

R_{11}, R_{12}, R_{13} → R_2 Motor Expression

R_{21} [Drawing the figure.]

R_{22} [Stating what is given. To be proved]



R_3 [Perception of known relationships in figure]

In △ ABD and CBD,

R_{31} ∠1 = ∠2 Given.

R_{32} ∠3 = ∠4 Given

R_4 [Recall in the light of analysis]

Conditions for equal triangles

$R_{41}(s - a - s)$

$R_{42}(a - s - a)$

$R_{43}(s - s - s)$

$R_3\begin{pmatrix} R_{31} \\ R_{32} \end{pmatrix}$, $R_4\begin{pmatrix} R_{41} \\ R_{42} \\ R_{43} \end{pmatrix}$ → R_5 [Which elements of R_3 and R_4 "fit in" with R_{11}?

Two pairs of equal angles, in R_3 and R_{42}] → R_6 [Which element, or elements in R_{42} differ from those in R_3? Included side] → R_7

R_7 [Does BD in △ABD equal BD in △CBD? Yes, "common."] → R_8 [Conclusion of equal triangles proved. (a s a) Thus, the common side, BD, is the unifying or link factor between the desired, and given and recalled facts]

Does this completion of the identity between the two series of facts result in a unified grouping, or classification, or "gestalt" of all the facts concerned? Furthermore, does recognition of success follow the solution of the exercise?

These analyses are illustrated in the exercise shown on p. 605.

These analyses led to the construction of a test prognostic of abilities in plane geometry. The situations and the responses to those situations are numerical and verbal, as well as spatial in content. The number and grouping of elements are to test the complex, organized abilities and certain of their constituent parts, in order to reveal the status of each student's abilities in perception, in recall, and in purposive thinking; they serve an additional purpose, also, that of differentiating between errors due to carelessness and to inabilities or low degrees of these abilities. The students' strengths and weaknesses thus detected may become the guides toward differentiated teaching emphases. The test includes elements in the form of directions to be followed, detecting of similar figures, recall, drawing and judging of conclusions, detection of correct reasons from classified series of statements for stated steps in geometric exercises, and motor expression of spatial imagination.

While this test was being constructed, another test appeared—"Orleans Geometry Prognosis Test." Although the purpose of each is identical, the philosophy underlying the construction of each and the nature of the content are strongly contrasting. The philosophy of its authors (Orleans) seems to be that prognosis of student abilities to learn geometry is best secured by student responses to varied samplings of geometric situations; *i.e.*, that prognosis of how students will learn geometry is by means of how they do learn when responding to actual geometric situations. On the contrary, the philosophy underlying the test (Perry) first described is that the more efficient prognosis will be effected by means of tapping the students' potential abilities which, when directed, can lead to effective mastery of plane geometry The "Orleans Geometry Prognosis Test" requires: Detection of identical meanings, of named angles, of angle size, of complementary and supplementary angles, of relative angle and side positions, of equal angles and of equal sides and double lengths of line-segments; perception of statement meanings due to perception of marked equal parts; mere detection of "if" and "then" clauses; and purposive thinking due to the selection of correct reasons from a given series of unclassified statements.

Both tests were given, though in reverse order, to students in two classes in plane geometry in Lincoln High School during the second semester of 1929–1930. Since the papers of students who were repeating the course or who were absentees on one day could not be included, there were forty-two students whose responses to the elements of each test were studied in detail and comparatively. In addition, there were two methods used for measuring the achievement of these students following their study of Book I: One was the study of student responses to Test 1 of "Hart's Geometry Tests," a test which is a part of the regular class work in this school; the other means were the grades assigned by the class instructors, although the teachers' estimates of student achievement were influenced to some extent by the responses of these students to the Hart test. Students' relative positions according to their total scores on each prognostic test, also their IQ's,* were charted with their achievement positions indicated by total scores on the Hart test and by instructors' grades Two techniques, that of the Pearson Product Moment Method and that of the Pearson Coefficient of Mean Square Contingency, were used to measure the relative amounts of correlation detected by each of these three methods of prognosis and two methods of determining achievement. The total r's are stated in the first part of Table I, the partial r's, with IQ's "partialled out," in the second part

TABLE I

	Total r's between methods of determining prognosis and achievement in geometry			Partial r's, (r's from which IQ's have been partialled out)	
	Perry	Orleans	IQ	Perry	Orleans
Grade	.61	66	48	49	50
Hart	63	78	35	.56	42
IQ .	50	76			

From these tables, and especially from the latter part of Table I, it appears that Orleans' test draws the more heavily upon intelligence as estimated by the Terman Group Test of Mental Ability This is noted from the drop of .12 in the measurement of the correlation, when IQ's have been partialled out, between the Perry test and the teachers' grades, and similarly a drop of .16 between Orleans' test and teachers' grades. When achievement is determined by the scores on the Hart test, there is a loss of .07 in the partial "*r*" from the total "*r*" between scores on the Perry and on the Hart tests, but similarly a much larger loss of .36 between scores on the Orleans and Hart tests. Furthermore, since the partial *r*'s (.49 and .50) are nearly identical, but the partial *r* (.56) is much larger than the partial *r* (.42), the Perry and the Orleans tests seem to prophesy achievement equally well *if* achievement is estimated by teachers' grades, but prognosis from scores on the Perry test is the more efficient *if* achievement is measured by scores on the Hart test. (Student time required for responding to the Perry and to the Orleans tests was forty-five, and seventy minutes, respectively.)

---

* The IQ's, obtained from the files in the office of the high school, were determined by the Terman Group Test of Mental Ability

As the number of cases is small, further analysis of correlation, as detected by three methods of prognosis and two methods of achievement, was accomplished by means of the Pearson Coefficient of Mean Square Contingency. Approximately equal proportions of the range of scores on the Perry and the Orleans tests, and those parts of the range of IQ's as used in the differentiation of class members, were compared with the teachers' grades,* as interpreted by the school administrators, and with the published quartiles of scores on the Hart test; these appear in Table II.

TABLE II.—STUDENT PLACEMENT ACCORDING TO PROGNOSIS AND ACHIEVEMENT

Test	Scores or IQ's	Teacher's grades				Hart scores			
		7	6–5	4–3	2–1	27–33	34–40	41–45	46–50
Perry	62–77	.		11	10		1	8	12
	50–61	2	5	6	1	1	4	6	3
	42–28	1	2	4		1	4	1	1
Orleans	80–140	1	2	14	9		2	10	14
	40–79	1	3	6	2	1	4	5	2
	8–39	1	2	1		1	3	.	.
IQ'S	120–140	.	.	3	5			2	6
	95–110	2	3	18	6		6	13	10
	70–94	1	4			2	3		.

---

* The grades, (1 and 2, 3 and 4, 5 and 6, and 7) are interpreted in per cents (in the nineties, eighties, seventies and below seventy, or "failure"), respectively

The Pearson Coefficients of Mean Square Contingency,* computed from Table II, are stated in Table IIa (The amounts of these coefficients could not exceed .866 *)

TABLE IIa - C_2's BETWEEN PROGNOSIS AND ACHIEVEMENT

	Perry	Orleans	IQ
Grade	55	43	61
Hart	51	57	63

The differences between the C_2's of the Perry scores and the measures of achievement, and the corresponding C_2's of the Orleans scores and of the IQ's, range from — 12 through .06 to .12. Thus, although there is association between each method of prognosis and of achievement, no marked superiority in their measurement was shown by any one of the techniques used

In conclusion: Prognosis of abilities to solve exercises in geometry seems the more efficient when based upon analysis of the requisite abilities and their constituent parts This means of prognosis (Perry test) not only requires less student time but it also duplicates less those abilities interpreted by the IQ's, than does the Orleans test. Scores on the Perry Geometry Prognosis Test are proving indicative of student status in achievement, and furthermore of information concerning student strengths and weaknesses information of paramount importance to the teacher who is to direct the learning of the students of plane geometry.

---

* The technique for computing these coefficients is described in "An Introduction to the Theory of Statistics" by G Udny Yule (pp. 66 67), also in "Statistics in Psychology and Education" by Henry E Garrett (pp 195 203)

THE DEVELOPMENT OF MENTAL ABILITY AT THE COLLEGE-ADULT LEVEL

MELVIN B. WRIGHT

University of Pennsylvania

If we may define behavior as the response of an organism to a stimulus situation, it is possible that a more complete knowledge of the organism, may enable us to make more accurate predictions, with respect to subsequent behavior. Experience has shown that there is a high degree of correlation, between the status of mental development and the quality of the personality products.

Most of the investigations of Mental Ability, have been confined to ages ranging from three to eighteen years This is partially due to lack of materials and methods for investigating the mental growth at the higher age levels. Furthermore, while it has been generally assumed that biological and environmental factors both contribute to mental development, but little progress has been made in analyzing the contributions of either.

The present investigation was undertaken, in order to determine what progress was taking place in mental development, at the College-Adult Level, an age range from eighteen to twenty-three years. Secondly, we propose to analyze this mental growth in such a way as to reveal the relationship between its biological aspects, and those which may be attributed, more directly to environmental pressures.

Terman has pointed out that though the Binet scale does not properly evaluate children who test at one hundred thirty, beyond the fifteen year age limit; this "does not mean that development ceases at this *time, but merely that the Stanford-Binet does not measure it.*"

Thorndike gives a parabolic curve, resembling the learning curve, rising rather abruptly from six to fifteen or sixteen years and then approaching a plateau which he describes as, "the probable form of the curve of intellect in relation to age."[2]

Thurstone has shown a mental growth curve, based on Binet test data, for ages three to fourteen years, concerning which he says, "A striking feature of this curve is that it continues to rise even at the age

---

[1] Terman, L M . "The Intelligence of School Children " P. 147

[2] Thorndike, E. L . "The Measurement of Intelligence " P 5

of fourteen years, with no indication of reaching a level. It certainly looks as tho the kind of intelligence which is measured by Binet tests and their variations, continues to grow as rapidly at fourteen years as it does at nine. This conclusion contradicts the statement frequently made, to the effect that test intelligence approaches an adult and more or less stable level at fourteen to sixteen years. The appearance of these curves, indicates that the growth of test intelligence continues beyond *the age of fourteen years . . . Common sense judgment* certainly favors the assumption that the average man of forty years is more intelligent than the average boy of twenty, but so far we have not been able to measure that difference. Instead of acknowledging this limitation in our measurement methods, we have not infrequently attempted to juggle with the definition of intelligence, to make it fit the measuring devices that are accessible."[1]

There are three curves in the graph to which Thurstone refers, all approximately parallel, one of which represents the mean ability of *successive age groups from three to fourteen years on Binet tests.* The second represents those who had a plus one sigma rating above the means of their respective age groups, and the third represents those who had a minus one sigma rating, below the mean for their respective age groups. All these curves show a correlation of better than .97 between test ability and age.

In a later article, Thurstone has extended this curve, by using similar data, for ages three to seventeen years, and interpolating the curve back to birth. The main part or central extension of the original curve, from three to fourteen years is slightly modified, so that *in the later curve, an inflection points develops between the ninth and* twelfth years, possibly about the eleventh year. From birth to the third year, the curve shows positive acceleration, which continues to the eleventh year, while from the fourteenth to the eighteenth year, the acceleration becomes about equally negative. This curve is shown in the diagram attached to this paper.

Thurstone concludes that the mental growth curve is asymptotic both to absolute zero and to an adult level. The lower limit is established at −.43 at birth, the upper limit is not ascertainable by computation. "These findings contradict the hypothetical mental growth curve of the text books, which is usually shown to have negative acceleration from birth without an inflection point."[1]

---

[1] Thurstone, L. L., A Method of Scaling Psychological Tests. *Journal of Educational Psychology*, Vol. XVI, No. 7, p. 442.

The present investigation attempts to measure the development of mental ability that lies beyond the range of Binet scales, as reported by Thurstone. It further seeks to ascertain whether mental development has reached the plateau stage between the ages of seventeen and twenty-three, at the College-Adult Level.

Thurstone regards his second mental growth curve, as giving a "better picture of continuity for mental growth regarded as a biological function." In pursuit of similar information, we have chosen a battery of tests, which we believe will give an index of the continuance or decline of this biological phase of mental growth. We have characterized this phase as Fundamental Ability.

We have also chosen a second battery of tests, which will evaluate another phase of mental growth, which may be defined as Complex Mental Processes. The former may be construed as concomitant with biological growth, while the latter represents the effect of educational and cultural pressures. The same tests were presented to all ages from eighteen to twenty-three on the assumption that there was equal competency within the group. This assumption was justified, in that all ages scored practically over the same ranges, in all tests, *with varying degrees of frequency.*

The results were obtained in the psychological laboratory at the University of Pennsylvania, from college men and women, during the three academic years, 1927–1928–1929–1930, as part of the required work in psychology courses. This testing program was instituted by Dr. Lightner Witmer, and has been developed over a number of years, by instructors in the clinical laboratory sections of first year psychology courses.

The tests, to which the students were subjected, were used as didactic experiments or demonstrations. They were administered under standard conditions and procedure, during the course of twelve class sessions, by a number of instructors, and scored on established standards. The significant measures for each test were derived from a combined protocol, of the various groups that had taken the tests. More than eleven hundred students have been tested in this manner during the three years, and have contributed to these results.

---

[1] Thurstone, L. L.: The Mental Growth Curve for Binet Tests. *Journal of Educational Psychology*, Vol. XX, No. 8, p. 569.

Dr. Brotemarkle, under whose supervision these courses were given, conceived the idea of using the scores to develop a "Mental Graph," by means of which various types of student ability might be portrayed graphically.[1] One of the distinctive features of this Mental Graph, is the analysis of Mental Ability into two parts, which he designates as: (1) Fundamental Abilities; and (2) Complex Mental Processes In evaluating any personality, or its products, it is manifestly of great importance, from both the diagnostic and prognostic points of view, to know whether an individual's inferiority or superiority, can be traced to either one or the other of these two aspects of his Mental Ability. Measurement on the Stanford-Binet, and other group or individual tests, affords no such insight

The pragmatic vindication of this procedure is evidenced by the types of graph actually obtained in practice, which Dr. Brotemarkle divides into the following seven classes.

1. Superior *meaning superior in Fundamental Ability and in* Complex Mental Processes
2. Superior *descending meaning superior in Fundamental Ability* descending in Complex Mental Processes
3. Median *ascending meaning median in Fundamental Ability* and ascending in Complex Mental Processes
4. Median *meaning median in Fundamental Ability and Complex* Mental Processes.
5. Median *descending meaning median in Fundamental Ability* and descending in Complex Mental Processes
6. Inferior *ascending meaning inferior in Fundamental Ability* and ascending in Complex Mental Processes
7. Inferior *meaning inferior in both Fundamental Ability and* Complex Mental Processes

It is by no *means inevitably assured that an individual with a* superior endowment of Fundamental Ability, will perforce remain superior in the field *of Complex Mental Processes* *Nor is it assured* that a median endowment of Fundamental Ability, may not rise or fall in the field of Complex Mental Processes *From these findings, it* became apparent that any estimate of the growth of mental ability at the College-Adult Level, must be analyzed, not only with respect to increasing age, but also with respect to the development of these two phases of mental ability.

---

[1] Brotemarkle, R A: College Student Personnel Problems *Journal of Applied Psychology*, Vol XI, No 6

Our investigation concerns the achievements of the various age groups from seventeen to twenty-three, comprising the College-Adult Level, in these two fields of Mental Ability. In forming the age groups, each one was made to include all individuals who had passed a given birthday, but had not passed the age level of the next higher group. Thus the eighteen year old group comprises those who have passed their eighteenth birthday, but have not yet reached the nineteenth.

The results obtained by the seventeen year old group, are not to be regarded as highly significant, owing to the very few cases available at this age level. The twenty-three plus group, represents a small miscellaneous company, ranging from twenty-three years to as high as 40 years. Their heterogeneity and limited number place a high degree of unreliability on the results obtained by them. These facts must be constantly borne in mind, when reading the extremities of our subsequent tabulations.

More than eleven hundred students were tested, though not all of these took all of the tests, due to absences and interruptions of the laboratory sessions. From this mass of material, we had available, between six hundred and seven hundred scores for each test, about equally divided between men and women. The scores were grouped in respective protocols for each age level A subsequent investigation of these results will be conducted, on the basis of sex, to ascertain what differences may prevail between them.

Wherever we had a score, belonging to an individual of a particular age group, it was recorded, regardless of how many of the twelve tests, the individual may have missed Our final results are therefore more significant, from the point of view of achievement in relation to age, since the results of no two tests, include scores by exactly and exclusively the same individuals. Nearly all the students took almost all of the tests, but some who took test number two, for example, may have missed tests three and five and *vice versa*. The results therefore, are based on scores, belonging to certain age groups, rather than to certain groups of persons of a given age. This factor dispels, to some extent, the possibility of a fortuitous selection of subjects Ours is a random sample of the College-Adult Population.

Following Brotemarkle's analysis of Mental Ability into the component parts of Fundamental Ability and Complex Mental Processes, the six tests which were used to measure each one of these factors, in this investigation, are listed below.

FUNDAMENTAL ABILITIES, NAMES OF TESTS	COMPLEX MENTAL PROCESSES, NAMES OF TESTS
1 Memory Span for Digits	1 Opposites (Roback)
2. Memory Span for Syllables	2 Abstraction (Roback)
3. Memory Span for Ideas	3 Reference (Roback)
4. Taylor Number Test	4 Definition (Brotemarkle)
5. Aussage (Undirected Attention)	5 Judgment (Roback)
6. *Ausfrage (Directed Attention)*	6 *Trabue Language Scale J*

These tests have been presented individually by their authors and subjected to analysis by Brotemarkle[1] and Miller[2] It is sufficient to indicate their general characteristics and the differences between the two batteries

The battery used for evaluating Fundamental Abilities, seeks to ascertain the relative plasticity of the organized protoplasm, its impressionability. The memory span tests are almost as simple as the patellar reflex, but *limits of achievement vary with the individual*, as does the excursion of the foot in the knee jerk The remaining three in this battery, are used to explore the field of attention. Is the individual *alert and receptive to the world about him?* To what extent does the world intrude itself upon his personality, and make impressions upon him. These three offer another clue to the individual's plasticity, his impressionability, the ease or difficulty with which he can be modified or adjusted to his environment.

The tests for Complex Mental Processes, investigate a different aspect of mental function in personality Here we observe and seek to evaluate, the influence and results of training and environment, the height of the cultural level, the clearness or obscurity of thought processes, the power of abstract thought, the ability to construct and compare concepts, keenness of discrimination, the precision with which one may describe or define, and so on. The whole range of complex mental functions in personality invites our inquiry, but at best, we can only select a few samples in the directions indicated above.

---

[1] Brotemarkle, R A.: College Student Personnel Problems. *Journal of Applied Psychology*, Vol XI, No 6.

[2] Miller, Karl G : The Competency of Fifty College Students *The Psychological Clinic*, Vol XIV, Nos. 1–2

We have treated the results obtained by these two batteries, separately. The scores for each test for each age group, were tabulated and the average and sigma determined A composite score for each age group, for each battery, was then developed. These composite scores are known as the Fundamental Ability Index and the Complex Mental Processes Index.

The method of obtaining these composite scores consisted first of all in weighting each of the six tests of each battery separately. The amount by which each test should be weighted was determined by the ratio, which its sigma bore, to the average sigma of the six tests in the battery. That is, the average sigma of the six tests, was divided by each of its constituent sigmas, and the resulting quotient was the weight chosen for that test. The mean of each test was multiplied by its respective multiplier; the results summated, and an average mean for the six tests derived, forming a composite score or index.

We have designated this statistical procedure as the average sigma method of weighting or combining test scores. Each of the six constituent elements of the composite mean, bears the same mathematical relation to the original from which it was derived, that its respective sigma bears to the average sigma of the battery.

This method, offered a consistent scheme for treating many combinations of scores in both batteries, and further eliminated the possibility that obtained results might have been influenced by a fortuitous choice of weights for the various tests. These indices were then subjected to the statistical analysis indicated in the discussion of results. Correlations were obtained by the Pearson Product Moment Formula.

A method of combining test scores, somewhat similar to that used in this investigation, is described by Garrett.[1] The "Chance R" for multiple Correlation, was obtained by the following formula:[2]

$$\text{"Chance } R\text{"} = \frac{\sqrt{n-1}}{\sqrt{N}}$$

ANALYSIS OF RESULTS

The means attained by the various age groups in the tests which constitute the battery for Fundamental Abilities, are found in Table I

---

[1] Garrett, H. E.: "Statistics in Psychology and Education." P 280

[2] Garrett, H. E.: "Statistics in Psychology and Education." P. 239

In as much as the populations of the various age groups differed considerably, the greater proportion of mathematical reliability belongs to the group having the largest population

The scores of the seventeen year age group, have been discarded in subsequent computations, because their small number gives an exceedingly high degree of unreliability to these scores. Furthermore, they represent a highly selected group of individuals, who by reason of an early start, special teaching, ability to skip a grade, or some other reason, reached the University before their eighteenth birthday. The size of this group as compared with the eighteen, nineteen, and twenty year group, indicates that some factor of selection was at work. If this factor happened to be superior mental ability, this marks them all the more as a highly selected group of seventeen year old individuals For our purposes we need random samples of all ages investigated, in sufficient numbers to give some degree of reliability.

The group listed as twenty-three plus, represents miscellaneous ages as high as forty and is therefore not indicative of any age. The scores of the twenty-two year old group have been used, though they number but twice the size of the seventeen year old group. However, they do not indicate that special factors contributed to their selection, which were not also applicable to ages eighteen to twenty-one. Nevertheless, due consideration must be given their small population in evaluating their indices.

TABLE I — SHOWING THE AVERAGE AND STANDARD DEVIATION FOR EACH AGE GROUP IN FUNDAMENTAL ABILITIES

Age	Frequency	MS digits		MS syllables		MS ideas		Undirected attention		Directed attention		Taylor number test	
		Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD
17	19	9 70	1 78	32 05	3 31	5 20	1 77	10 21	3 94	24 66	4 77	14 52	3 30
18	109	8 62	1 57	31 18	4 38	5 29	2 14	9 85	3 30	22 29	5 13	11 17	4 24
19	236	8 18	1 58	20 07	6 27	4 00	2 15	0 30	3 40	23 57	5 16	13 00	3 50
20	106	8 23	1 55	30 22	5 44	4 68	1 90	9 44	3 71	23 24	5 55	13 31	3 02
21	100	0 35	1 55	29 70	6 03	4 77	2 01	10 12	4 32	22 02	4 06	11 22	3 74
22	38	8 80	1 56	30 03	6 15	4 07	2 00	9 31	4 50	22 20	6 48	14 32	3 82
23+	12	8 57	1 45	30 13	4 48	5 20	1 07	9 01	3 22	21 20	6 00	14 02	4 16

Table I gives little indication of the relative abilities of the respective age groups, by observation.

TABLE II.—SHOWING THE AVERAGE AND STANDARD DEVIATION FOR EACH AGE GROUP IN COMPLEX MENTAL PROCESSES

Age	Frequency	Opposites		Abstraction		Reference		Judgment		Trabue		Definition	
		Average	SD	Average	SD	Average	SD	Average	SD	Average	SD	Average	SD
17	18	49 88	10 65	51 05	12 65	50 82	4 62	60 50	12 90	11 00	1 67	47 36	12 90
18	98	49 67	12 85	51 74	14 35	28 68	7 02	57 49	14 50	9 52	1 78	48 53	15 05
19	226	50 00	14 15	47 50	15 70	29 65	5 97	55 81	12 75	9 50	1 61	43 32	15 65
20	185	48 50	13 35	51 96	15 35	28 50	5 49	57 42	11 40	9 49	1 64	46 10	14 85
21	93	50 73	13 35	55 06	16 30	29 08	5 82	55 00	11 40	9 24	1 59	45 79	10 35
22	37	46 66	15 20	54 22	21 40	29 62	6 64	53 68	12 65	9 60	1 67	46 34	12 85
23+	42	49 71	11 50	51 29	13 35	26 96	6 50	52 90	13 15	9 52	1 46	45 33	14 45

Table II shows in similar manner, the means achieved by the various age groups in the tests comprising the battery for Complex Mental Processes. The results do not greatly facilitate the discovery of trends in mean ability or variability, by observation

TABLE III.—SHOWING THE COMPOSITE AVERAGE SCORE FOR EACH AGE GROUP IN FUNDAMENTAL ABILITIES

Age	Frequency	Mean
18	109	15 14
19	236	14 37
20	190	14 20
21	100	15 14
22	38	15 25

Table III shows the results of combining the six Fundamental Ability test scores, into a composite or index, for each group. The eighteen and nineteen year old groups appear to be superior to the twenty year old group, and the twenty-one and twenty-two year old groups are also superior to the twenty year old group. The average age of this entire College-Adult Level, weighted according to the number of each population, is 19.59 years, while the average mean for the entire group is 14 63 in Fundamental Ability. The mean age of this group of College adults therefore lies between nineteen and twenty years, while the mean ability of the college adult is represented by a score that is closer to the score of these two ages, than it is to that of any other age group.

TABLE IV.—SHOWING THE COMPOSITE SCORE FOR EACH GROUP IN COMPLEX MENTAL PROCESSES

Age	Frequency	Mean
18	98	43 84
19	226	45 38
20	185	46 35
21	93	47 20
22	33	47 09

The results of combined scores for the Complex Mental Processes are shown in Table IV. Here the trend toward increasing ability with increasing age is more readily observed. The size of the means for Complex Mental Processes as compared with the size of the means for Fundamental Ability is not indicative, since these sizes represent the arbitrary numerical values of the respective scales by which the tests were graded.

The correlation of Fundamental Abilities with age is + 17 with a PE of + 025. This represents a low degree of correlation, yet when portrayed graphically, it reveals some significance The correlation of Complex Mental Processes with age is + 95, with a PE of + 003. The Multiple correlation of age with (Fundamental Ability and Complex Mental Processes) is + 97 with a "Chance R" of + 056.

The comparison of the respective trends indicates, first of all, that these two batteries have tapped two different kinds or phases of ability. One of these we have defined as native endowment, Fundamental Ability, or mental growth considered as a biological function. The other is defined as a highly complex intellectual response to educational and cultural pressures; i e. the development of Complex Mental Processes From this point of view, mental growth at the higher intellectual levels becomes less of a biological function and more of a response to selected environmental pressures

It must be recalled that we are dealing with a chronological period ranging from eighteen through twenty-two years, hence what may be true of these particular ages, may not be true above or below them. Until now, the correlation between mental growth and age for this period has been assumed as zero, which would give us a plateau on the curve. By actual measurement, of the correlation, we have found it to be +.17, which gives a gradual incline upward toward maturity.

The trend of Complex Mental Processes ability, rises from eighteen to twenty-two years, with a slight drop of 11 at this last stage, possibly due to the small population at this point. Increasing age, lived under the stimulus of cultural and educational environment, has provided a more extensive orientation, a richer background from which all mental function products may be derived. The fruits of added years of experience are manifest in the trend toward greater ability to make highly complex responses with increasing age

In order to present these correlations graphically, the slopes of the regression lines were determined by the formula:

$$M_{\frac{b}{a}} = \frac{\Sigma xy}{\Sigma x^2}$$

where M represents the slope of the regression of (1) Fundamental Ability on age and (2) of Complex Mental Processes on age. b represents the mean of either array of ability indices, and a represents the array of ages. x and y represent deviations from the means of age and ability, respectively. The sum of these deviations is determined by the size of the populations in each group

The regression of Fundamental Ability on age gave a coefficient of +.06 and the regression of Complex Mental Processes on age gave a coefficient of +.94. Further reduction of these values to sigma or standard units, facilitated their representation on the same graph. The formula in either case became:

$$M_{\frac{b}{a_s}} = \frac{M_{\frac{b}{a}}}{\sigma_b}$$

where s means "standard units" or "sigma units" and σ_b is the sigma of the array of either ability taken separately. This gave the slope of the regression lines in terms of standard units, that for Fundamental Ability became $y = .15x$; and that for Complex Mental Processes, $y = .93x$.

The Standard Error for the Coefficient of Regression is obtained by the formula:

$$\sigma_{y_{b/a}} = \frac{\sigma_b}{\sigma_a}\sqrt{\frac{1-r^2}{N}},$$

where σ_a is the sigma of the age array; r is the coefficient of correlation between ability and age, and N the number of cases.

The Standard Error of the Coefficient of Regression, of Fundamental Ability on age is found to be +.013, and that for Complex Mental Processes on age is found to be + 012.

The values indicate first of all, the high degree of improbability that these two curves (1) and (2) can ever change sufficiently to move with the same degree of acceleration, within the range of eighteen to twenty-three years, thus furnishing added assurance, in terms of increased probability, to any conclusions justified by these curves.

The limits set upon the slope of the Fundamental Ability curve (2) by its standard of error, indicate that there is but very slight probability that the mental growth curve will reach a plateau during this period. In other words, there is exceedingly high probability, that this portion of the curve will always rise slightly, and furthermore, that its present location is the most probable.

In a similar manner, the probability that the Complex Mental Process curve (1) will ever reach a plateau during this period is even more remote, with the strong presumption that its present position is the most likely, for any random sampling at the College-Adult Level.

Diagram I is a graphic presentation of these results Our problem is to reconcile these two different aspects of mental growth. *It is evident from Section 2 of the curve, that the hitherto assumed plateau was relatively though not absolutely correct.* That the curve is approaching a plateau at age seventeen is certain, but that this plateau lies beyond twenty-three is also certain. Fundamental Ability, native endowment, is still slightly on the increase through the twenty-second year.

Section 1 of the dual curve, resembles very much the curves obtained for test ability correlated with ages eight to fifteen on the National Intelligence Test; also that from eighteen months to six and one half years on the Stutsman Test, and that from eleven to sixteen years on the Otis Advanced Intelligence Test.[1] The rate of its acceleration is in sharp contrast with that of Section 2. The relative height above Section 2 is insignificant, as this position was arbitrarily chosen with respect to other features of the dual curve. The relative slope of these two curves is the significant factor.

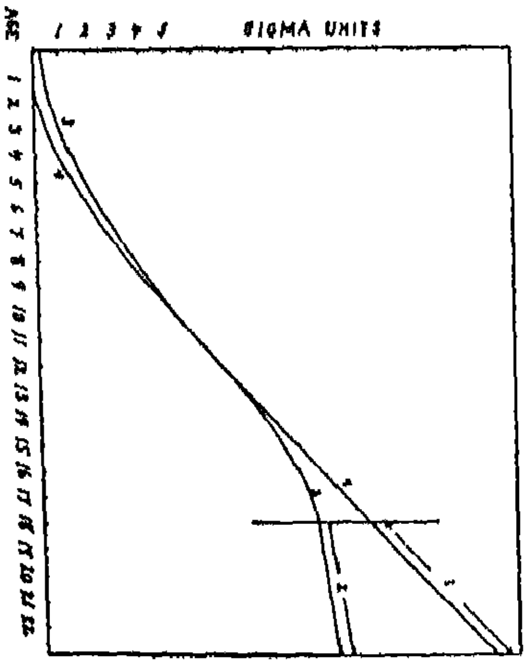


Diagram I

Curves showing the relative slope of the Regression Lines, from Ages eighteen to twenty-three.

1. Regression of Complex Mental Processes on Age, eighteen to twenty-three
2. Regression of Fundamental Abilities on Age, eighteen to twenty-three
3. Thurstone Mental Growth Curve from zero to seventeen years
4. Theoretical Curve of Complex Mental Process Development, zero to seventeen.

It would appear from the shape of Section 2, that the native endowment or mental growth regarded as a biological function, is fast approaching its maximum between eighteen and twenty-three years.

---

[1] Thurstone, L. L.: The Absolute Zero in Intelligence Measurement. *Psychological Review*, Vol. 35, No. 3, May, 1928, p. 103.

On the other hand, the intellectual development and cultural status of this same group, continues to rise proportionately to their continued subjection to educational and cultural curricula. It seems reasonable to characterize Section 2 as indicative of rather elementary mental behavior, and Section 1 as indicative of rather highly complex mental behavior. We shall probably avoid some confusion in discussion by omitting the doubtful inference to "intelligence."

A further characterization, which differentiates these two types of mental growth, might regard Section 2 as the evolution of the inherent qualities of the protoplasm a simple behavioristic estimate of mental growth. The other curve, Section 1, represents no denial of the former, but reveals a supplementary aspect of mental function in personality development, wherein the cultural and educational influences of society are manifest. This curve shows a much higher correlation with chronological age. Since these two types of behavior are produced by the same biological organism each as inevitable as the other —it would seem that every stimulus situation offered an opportunity for selective reaction or choice of mental function. We may react biologically, at the animal level, or we may react as intellectual personalities at a higher level. Perhaps the greater degree of employability, tends toward the higher rather than the lower level, at the College-Adult Level, especially if we have regard for the present value of correlation coefficients.

Our next problem is to relate our results to previous investigations of mental growth On the basis of the results obtained in this investigation, we have attempted to interpolate both Sections 1 and 2 of our dual curve, backward to birth.

We have accepted the mental growth curve of Thurstone, from three to seventeen years, as most indicative of this period. Without modification of either curve, our Fundamental Ability curve, Section 2, appears to be a continuation of his findings With some degree of confidence, the curve for mental growth, considered as a biological function, can now be drawn from birth to the twenty-third year. From the shape of Thurstone's curve (3), it is evident that the rate of this developmental process is not constant It has a positive acceleration to approximately the eleventh year, where it reaches an inflection point. There negative acceleration sets in with steadily increasing swiftness until we reach the seventeenth or eighteenth year, at which point it appears to enter upon a period of constancy through the twenty-second year. This type of mental growth increases as much or more in any single year, between seven and fourteen years, as it does during the whole five year period from eighteen through twenty-two.

We have noted the similarity between our Complex Mental Process curve, Section 1, and the curves derived for other ages on the National and Otis tests. We have not endeavored to fit our curve to any of these as yet, though we propose to make this attempt in the subsequent investigation of our results on the basis of sex. For the moment, we have placed our Complex Mental Process curve, Section 1 on the chart in an arbitrary manner, preserving only its relative slope, and theoretically extending it toward zero ability. We have designed that it intercept the Thurstone curve at the eleven year point. This would bring zero ability for this type of mental ability at about the four year age, if the curve were carried through.

If however, the Complex Mental Process curve represents the development of a certain type of mental behavior, we should prefer to begin our theoretical curve, possibly at about eighteen months, with the inception of language ability in the individual, even though we may not be able to measure it here

Between its inception, and the sixth year, the Complex Mental Process curve (4) is almost equally accelerated with the Fundamental Abilitiy curve, though slightly below it. At the sixth year, the individual meets the educationally organized current of civilization, which is designed to shape and fashion his mental behavior He is subjected to educational, social and cultural pressures, represented by the school curriculum, his contacts with teachers, and his contemporaries in intellectual competition. Hence at this point, a more rapid acceleration of his Complex Mental Process ability takes place, which overtakes his Fundamental Abiltiy curve about the eleventh year. From this point to his twenty-third year, our assumed Complex Mental Process curve follows the trend indicated by our investigation.

This may account somewhat for the pronounced change that takes place in the individual's mental functioning just before the onset of puberty. According to our interpretive curve, at this point, the individual's behavior is determined less and less by pronounced biological mental response, and more and more by an intellectual or cultural response. This is in no sense a denial that our solution of many problems, even at upper age levels, rises no higher than biological mentality, but it is to say that we can and do develop potentialities that make reaction on a complex mental level, just as possible, and certainly more probable. What we speak of as "control," very *often means to delay a "biological mental" response, and refer the* solution of the problem to the higher level of Complex Mental Processes. *It is an observed phenomenon, that a pronounced change in* mental behavior characterizes the period from nine to twelve years. Possibly the wane of biological mental development and the continued increase of highly complex mental processes may partially account for this change. Our interpretive curves offer an interesting field in which to conduct further investigations

Our final consideration is the prediction of behavior. Frequent reference to our interpretive curves, will help to clarify the discussion. These curves have established the normal course of mental development. They predict that fifty per cent of the individuals of a given age will be found above and fifty per cent below them, in similar samples of the population. This may be regarded as a general diagnosis of mental behavior up to the twenty-third year.

Prognosis is simply the theoretical projection of diagnosis, and our interpretive curves, offer the most probable direction for prognostic projection. For example, since fifty per cent of the populations investigated, are either above or below the curve at three years of age and at twenty-three alike, it seems reasonable to assume that the group which was above at three years will also be the group which is still above at twenty-three, simply because the curve of normal mental development lies between these two groups all the way. Any other *assumption would deny the validity of the normal curve of mental* growth. It is mathematically impossible for a point one sigma below *the curve at three years, to reach a point one sigma above the curve* at twenty-three years, without tracing a new curve in its progress, which would be entirely different, than the curve which we have found to represent normal mental development

Diagnosis establishes the individual's relation to normality and measures his deviations from the normal mental growth curve, on a straight line, perpendicular with the chronological base line. Prognosis on the other hand, concerns the movement of this diagnostic deviation point, toward the right, in the direction in which the chronological base line extends, *but not parallel to it.* The Deviation point must move in a curve which is of the same general form as the normal mental growth curve. This is but another way of saying, what is widely recognized, that bright, normal and dull children develop mentally at approximately the same rate.

It is to be noted however that variability increases with age, hence positive and negative deviations increase slightly with age This fact helps to authenticate any prognosis based on accurate diagnosis, because it gradually separates the distance between the normal curve and deviate curves, with increasing years, thus making their coincidence or crossing even more remote

The shape of the Fundamental Ability curve, Sections 2 and 3, is an ardent protest against any prognosis of adult behavior, which is based on an infancy curve. The most thorough diagnosis of the one to five year period of this curve, gives no indication whatever of its direction between ten and fifteen or between eighteen and twenty-three years. Furthermore, the cultural and intellectual development, represented by the Complex Mental Process curve, Sections 1 and 4, by which the individual will be more constantly evaluated as a social unit, are so incipient from one to five years, that prognosis with respect to this phase of mental development, in this period, is very doubtful. In other words, the degree of acceleration with which mental development takes place with increasing age, *i.e.*, the shape of the normal development curves, can not be determined by prognosis. Now that these curves have been determined by investigation, prognosis must follow the course they indicate.

Brotemarkle has shown in his "Mental Graph Types" that the prediction of human behavior can only be based upon accurate diagnosis of an individual, with respect to the two phases of mental development indicated by our interpretive curves. A superior biological mental basis is not absolute assurance of superior ability with respect to the development of Complex Mental Processes

This is only to say that social non-conformity which reaches the acute stage of variability in criminal acts, is just as frequently the concomitant of normal biological mentality as it is of abnormal Conversely the normal biological mental endowment is quite capable of attaining high degrees of culture and social conformity. Biological mental development tells only part of the story, and prognoses made on this basis alone are subject to a very large probable error, large enough even to invalidate the prognoses

True prognosis must begin with accurate diagnosis, with respect to Fundamental Abilities and Complex Mental Processes, and must infer development in accordance with the curves. The fundamental endowment of the individual is basic, to be sure, but his ultimate value as a unit of society, is even more determined by the permutations and combinations of Fundamental Abilities, *i.e.*, the development of his Complex Mental Processes.

The following conclusions are indicated:

1. *Mental Behavior, as an index of mental growth at the College-Adult Level, should be analyzed into* Fundamental Abilities and Complex Mental Processes.

2. The development of Fundamental Ability, regarded as a biological function, has not yet reached the plateau stage between eighteen and twenty-three years.

3. Complex Mental Processes are more largely due to environmental pressures than are Fundamental Abilities, hence *they respond* and increase more rapidly at the College-Adult Level.

4. *Though we may not expect the* Complex Mental Processes to increase indefinitely, there is no evidence of an approaching plateau at age twenty-three.

5. Fundamental Abilities develop as much, or more, in any single year between seven and fourteen, as they do during the entire five year period from eighteen through twenty-two, at the College-Adult Level.

6. There is a possibility that the probable proportionate increase *of Complex Mental Process development, and the proportionate* decrease of Fundamental Ability development, just prior to or about the age of puberty, may partially account for the changes in mental behavior observable at this period.

7. While children up to eleven years of age may be more inevitably committed to reactions at the level of biological mental development, they become increasingly committed to reactions at the higher complex intellectual level thereafter.

8. Predictions of behavior should be based upon the shape of the *mental growth and behavior curves, which change conspicuously with* increasing years.

I wish to express my gratitude to Dr. George Gailey Chambers and Dr. H. M. Lufkin, of the Department of Mathematics, of the University of Pennsylvania, for criticism and suggestions in the development of statistical procedures; to Dr. Samuel W. Fernberger of the Department of Psychology of the University of Pennsylvania for *council and assistance in my investigation*, to Dr. Robert A. Brotemarkle, under whose supervision the investigation was carried out, for materials and suggestions which greatly facilitated the work, and to the instructors and students, whose cooperation over a period of three years made the materials available.

BIBLIOGRAPHY

Brooks, Fowler D.: "Changes in Mental Traits with Age." Teachers College Publications, Columbia University, 1921.

Brotemarkle, Robert A.: College Student Personnel Problems. *Journal of Applied Psychology*, Vol. XI, No. 6, 1927.

Cunningham, Kenneth S.: "The Measurement of Early Levels of Intelligence." Teachers College Publications, Columbia University, 1927.

Fernberger, S. W.: Statistical and Non-statistical Treatment of Results. *Psychological Clinic*, Vol. XIV, Nos. 3-4, May-June, 1922.

Foran, T. G.: The Constancy of the Intelligence Quotient. *Educational Research Bulletin*, Vol. I, No. 10, December, 1925.

Garrett, H. E.: "Statistics in Psychology and Education."

MacPhail, Andrew H.: "The Intelligence of College Students."

Miller, Karl G.: The Competency of Fifty College Students. *The Psychological Clinic*, Vol. XIV, Nos. 1-2, March-April, 1922.

Rosenow, Curt: Analysis of Mental Functions. *Psychological Monographs*, Vol. XXIV, No. 5, June, 1917.

Terman, L. M.: "The Intelligence of School Children."

Thorndike, E. L.: "The Measurement of Intelligence."

Thurstone, L. L.: A Method of Scaling Psychological Tests. *Journal of Educationa Psychology*, Vol. XVI, No. 7. Absolute Zero in Intelligence Measurement. *Psychological Review*, Vol. XXXV, No. 3. The Mental Growth Curve fo Binet Tests. *Journal of Educational Psychology*, Vol. XX, No. 8.

Tilton, J. W.: "The Relation between Association and the Higher Mental Processes." Teachers College Publications, Columbia University, 1926.

A MODIFICATION OF THE COMPUTATION OF THE MULTIPLE CORRELATION AND REGRESSION COEFFICIENTS BY THE TOLLEY AND EZEKIEL METHOD

AARON BAKST

Teachers College, Columbia University

The Tolley and Ezekiel method of handling the multiple correlation problem given in their paper[1] and further modified by Garrett[2] states the relation between the partial regression coefficients, b's, in the form of a set of linear normal equations that can be solved simultaneously.

These equations are given in two forms: either in terms of the mean product sums, or in the terms of the correlation coefficients defined by the relations:

$$p_{12} = r_{12}\sigma_1\sigma_2, \quad r_{12} = \frac{\Sigma(x_1x_2)}{N\sigma_1\sigma_2}, \quad p_{12} = \frac{\Sigma(x_1x_2)}{N}$$

Rewriting Garrett's equations (3) (p. 42) we have:

$$\begin{aligned}
r_{12} &= A_{21}b_{12.34\ldots n} + r_{23}A_{31}b_{13.24\ldots n} + r_{24}A_{41}b_{14.23\ldots n} + \cdots + r_{2n}A_{n1}b_{1n.23\ldots} \\
r_{13} &= r_{23}A_{21}b_{12.34\ldots n} + A_{31}b_{13.24\ldots n} + r_{34}A_{41}b_{14.23\ldots n} + \cdots + r_{3n}A_{n1}b_{1n.23\ldots} \\
r_{14} &= r_{24}A_{21}b_{12.34\ldots n} + r_{34}A_{31}b_{13.24\ldots n} + A_{41}b_{14.23\ldots n} + \cdots + r_{4n}A_{n1}b_{1n.23\ldots}, \\
&\vdots \\
r_{1n} &= r_{2n}A_{21}b_{12.34\ldots n} + r_{3n}A_{31}b_{13.24\ldots n} + r_{4n}A_{41}b_{14.23\ldots n} + \cdots + A_{n1}b_{1n.23\ldots},
\end{aligned}$$

where the values of A_{21}, A_{31}, A_{41}, etc. are represented by σ_2/σ_1, σ_3/σ_1, σ_4/σ_1, etc., respectively.

---

[1] Tolley, H. R. and Ezekiel, M. J. B.: A Method of Handling Multiple Correlation Problems. *Journal of American Statistical Association*, Vol. XVIII, Dec., 1923, pp. 993-1003.

[2] Garrett, Henry E., A Modification of Tolley and Ezekiel's Method of Handling Multiple Correlation Problems. *Journal of Educational Psychology*, Vol. XIX, Jan., 1928, pp. 15-19.

The value of $b_{12 \cdot 34 \ldots n}$ can be written in determinantal form as:

$$b_{12 \cdot 34 \ldots n} = \frac{\begin{vmatrix} r_{12} & r_{23}A_{33} & r_{24}A_{44} & \ldots & r_{2n}A_{nn} \\ r_{13} & A_{33} & r_{34}A_{44} & \ldots & r_{3n}A_{nn} \\ r_{14} & r_{34}A_{33} & A_{44} & \ldots & r_{4n}A_{nn} \\ \vdots & & & & \vdots \\ r_{1n} & r_{3n}A_{33} & r_{4n}A_{44} & \ldots & A_{nn} \end{vmatrix}}{\begin{vmatrix} A_{22} & r_{23}A_{33} & r_{24}A_{44} & \ldots & r_{2n}A_{nn} \\ r_{23}A_{22} & A_{33} & r_{34}A_{44} & \ldots & r_{3n}A_{nn} \\ r_{24}A_{22} & r_{34}A_{33} & A_{44} & \ldots & r_{4n}A_{nn} \\ \vdots & & & & \vdots \\ r_{2n}A_{22} & r_{3n}A_{33} & r_{4n}A_{44} & \ldots & A_{nn} \end{vmatrix}}$$

The columns of both determinants of the numerator and denominator have common factors A_{33}, A_{44}, A_{55}, etc. that can be taken outside of the determinant. Moreover, the factors of the determinant of the numerator appear as factors in the determinant of the denominator. Therefore, the value of $b_{12 \cdot 34 \ldots n}$ can be rewritten as:

$$b_{12 \cdot 34 \ldots n} = \frac{\begin{vmatrix} r_{12} & r_{23} & r_{24} & \ldots & r_{2n} \\ r_{13} & 1 & r_{34} & \ldots & r_{3n} \\ r_{14} & r_{34} & 1 & \ldots & r_{4n} \\ \vdots & & & & \vdots \\ r_{1n} & r_{3n} & r_{4n} & \ldots & 1 \end{vmatrix}}{A_{21}\begin{vmatrix} 1 & r_{23} & r_{24} & \ldots & r_{2n} \\ r_{23} & 1 & r_{34} & \ldots & r_{3n} \\ r_{24} & r_{34} & 1 & \ldots & r_{4n} \\ \vdots & & \vdots & & \vdots \\ r_{2n} & r_{3n} & r_{4n} & \ldots & 1 \end{vmatrix}}$$

In general the value of $b_{1k \cdot 23 \ldots n}$ can be expressed in this form, and noticing that in the determinant of the numerator we replace with the column r_{12}, r_{13}, r_{14}, etc. the column in which the coefficients of $b_{1k.23\,.\,.\,n}$ appear, and moreover that transposing the k-th column in the position of the first column and the k-th row in the position of the first row of both determinants we do now change the value of $b_{1k.23\,.\,.\,n}$ and then replace the k-th column in the numerator in the same manner as in case of obtaining the expression of $b_{12.23\,.\,.\,n}$ the expression of $b_{1k.23\,.\,.\,n}$ is as follows:

$$b_{1k.23\,.\,.\,n} = \frac{\sigma_1}{\sigma_k} \frac{\begin{vmatrix} r_{1k} & r_{2k} & r_{3k} & \cdots & r_{kn} \\ r_{12} & 1 & r_{23} & \cdots & r_{2n} \\ r_{13} & r_{23} & 1 & \cdots & r_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ r_{1n} & r_{2n} & r_{3n} & \cdots & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{2k} & r_{3k} & \cdots & r_{kn} \\ r_{2k} & 1 & r_{23} & \cdots & r_{2n} \\ r_{3k} & r_{23} & 1 & \cdots & r_{3n} \\ \cdot & \cdot & \cdot & & \cdot \\ r_{kn} & r_{2n} & r_{3n} & \cdots & 1 \end{vmatrix}} \tag{2}$$

The evaluation of these two determinants, which are identical, with the exception of their first columns can be done as follows:

Consider a determinant

$$D = \begin{vmatrix} a_1 & b_1 & c_1 & d_1 & \cdots & k_1 & l_1 \\ a_2 & b_2 & c_2 & d_2 & \cdots & k_2 & l_2 \\ a_3 & b_3 & c_3 & d_3 & \cdots & k_3 & l_3 \\ \cdot & & & & & & \cdot \\ a_n & b_n & c_n & d_n & \cdots & k_n & l_n \end{vmatrix}$$

Multiplying each term of every column by the first term of the column on the right, with its sign changed, and adding to it the product of the corresponding terms in the same row in the column on the right with the first term of the original column, we have:

$$(-1)^{n-1}b_1c_1d_1 \cdots k_1l_1D = \begin{vmatrix} 0 & 0 & \cdots & 0 & l_1 \\ -a_2b_1+a_1b_2 & -b_2c_1+b_1c_2 & \cdot & -k_2l_1+k_1l_2 & l_2 \\ -a_3b_1+a_1b_3 & -b_3c_1+b_1c_3 & & -k_3l_1+k_1l_3 & l_3 \\ \vdots & & & & \vdots \\ -a_nb_1+a_1b_n & -b_nc_1+b_1c_n & & -k_nl_1+k_1l_n & l_n \end{vmatrix}$$

and this by the expansion of the determinant by minors (taking them along the first row) is transformed into:

$$(-1)^{n-1}b_1c_1d_1 \cdots k_1l_1D = (-1)^{n-1}l_1 \begin{vmatrix} -a_2b_1+a_1b_2 & -b_2c_1+b_1c_2 & \cdot & -k_2l_1+k_1l_2 \\ -a_3b_1+a_1b_3 & -b_3c_1+b_1c_3 & & -k_3l_1+k_1l_3 \\ \vdots & & & \vdots \\ -a_nb_1+a_1b_n & -b_nc_1+b_1c_n & \cdot & -k_nl_1+k_1l_n \end{vmatrix}$$

or finally

$$D = \frac{1}{b_1c_1d_1 \ldots k_1} \begin{vmatrix} \begin{vmatrix} a_1b_1 \\ a_2b_2 \end{vmatrix} & \begin{vmatrix} b_1c_1 \\ b_2c_2 \end{vmatrix} & \begin{vmatrix} c_1d_1 \\ c_2d_2 \end{vmatrix} & \ldots & \begin{vmatrix} k_1l_1 \\ k_2l_2 \end{vmatrix} \\ \begin{vmatrix} a_1b_1 \\ a_3b_3 \end{vmatrix} & \begin{vmatrix} b_1c_1 \\ b_3c_3 \end{vmatrix} & \begin{vmatrix} c_1d_1 \\ c_3d_3 \end{vmatrix} & \ldots & \begin{vmatrix} k_1l_1 \\ k_3l_3 \end{vmatrix} \\ \vdots & & & & \vdots \\ \begin{vmatrix} a_1b_1 \\ a_nb_n \end{vmatrix} & \begin{vmatrix} b_1c_1 \\ b_nc_n \end{vmatrix} & \begin{vmatrix} c_1d_1 \\ c_nd_n \end{vmatrix} & \ldots & \begin{vmatrix} k_1l_1 \\ k_nl_n \end{vmatrix} \end{vmatrix} \qquad (A)$$

Using this transformation, and noticing that the terms taken out in the fraction on the left of the determinant (A) will be equal for both determinants of (2) the expression for $b_{12.34\ldots n}$ can be rewritten as follows:

$$b_{1k.23\ldots n}=\frac{\sigma_1}{\sigma_k}\,\frac{\begin{vmatrix}\begin{vmatrix}r_{1k}&r_{2k}\\r_{12}&1\end{vmatrix}&\begin{vmatrix}r_{2k}&r_{3k}\\1&r_{23}\end{vmatrix}&\cdots\\\begin{vmatrix}r_{1k}&r_{3k}\\r_{13}&r_{23}\end{vmatrix}&\begin{vmatrix}r_{2k}&r_{3k}\\r_{23}&1\end{vmatrix}&\cdots\\\vdots&&\\&&\cdots\end{vmatrix}}{\begin{vmatrix}\begin{vmatrix}1&r_{2k}\\r_{2k}&1\end{vmatrix}&\begin{vmatrix}r_{2k}&r_{3k}\\1&r_{23}\end{vmatrix}&\cdots\\\begin{vmatrix}1&r_{3k}\\r_{3k}&r_{23}\end{vmatrix}&\begin{vmatrix}r_{2k}&r_{3k}\\r_{23}&1\end{vmatrix}&\cdots\\\vdots&&\\&&\cdots\end{vmatrix}}$$

Continuing the process of transforming the determinants in the same manner the complete evaluation of the $b_{1k.23\ldots n}$ is concluded.

The advantage of this method of evaluation lies in the fact that in the case of each particular regression coefficient only one column is inserted in a new position in the determinant, and only one row and column are transposed in the determinants of the numerator and denominator Therefore, the evaluation of any subsequent regression coefficients is simplified, for as it is seen from the nature of determinant (A) certain results are retained throughout the entire process of computing.

It can be readily seen that this process of computing offers another advantage, namely that any regression coefficients of orders less than "n" *can be obtained during the computation of* $b_{1j.23\ldots n}$.

Let us write down the set of equations predicting $b_{1k.23}$

$$\begin{aligned}r_{12}&=A_{21}b_{12.3k}+r_{23}A_{31}b_{13.2k}+r_{2k}A_{k1}b_{1k.23}\\r_{13}&=r_{23}A_{21}b_{12.3k}+A_{31}b_{13.2k}+r_{2k}A_{k1}b_{1k.23}\\r_{1k}&=r_{2k}A_{21}b_{12.3k}+r_{3k}A_{31}b_{13.2k}+A_{k1}b_{1k.23}\end{aligned}$$

The reader can readily construct the value of $b_{1k.23}$ in the form of determinants, it is (following the process described above and transposing rows and columns)

$$C_{11\,23}\frac{\sigma_1}{\sigma_k}\frac{\begin{vmatrix} r_{1k} & r_{2k} & r_{3k} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{2k} & r_{3k} \\ r_{2k} & 1 & r_{23} \\ r_{3k} & r_{23} & 1 \end{vmatrix}}$$

It may be observed that when solving the original equations for $b_{1k.23\ldots n}$ the values of $b_{1k.2}$, $b_{1k.23}$, $b_{1k.234}$, etc., may be computed.

The formula for the coefficient of multiple correlation, R, as given by Garrett (p. 48, *formula* 7) is

$$R^2_{1(23\ldots n)} = \frac{b_{12.34\ldots n}r_{12}\sigma_2 + b_{13.24\ldots n}r_{13}\sigma_3 + \cdots + b_{1n.23\ldots(n-1)}r_{1n}\sigma_n}{\sigma_1^2}$$

substituting in this formula the values of $b_{1k.23\ldots n}$ from formula (2) we have

$$R^2_{1(23\ldots n)} = b_{12.34\ldots n}r_{12}A_{21} + b_{13.24\ldots n}r_{13}A_{31} + \cdots + b_{1n.23\ldots(n-1)}r_{1n}A_{n1}$$

or in other form

$$R^2_{1(23\ldots n)} = \sum_{k=2}^{n} r_{1k}\frac{\begin{vmatrix} r_{1k} & r_{2k} & r_{3k} & \ldots & r_{kn} \\ r_{12} & 1 & r_{23} & \ldots & r_{2n} \\ r_{13} & r_{23} & 1 & \ldots & r_{3n} \\ \vdots & & & & \\ r_{1n} & r_{2n} & r_{3n} & \ldots & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{2k} & r_{3k} & \ldots & r_{kn} \\ r_{2k} & 1 & r_{23} & \ldots & r_{2n} \\ r_{3k} & r_{23} & 1 & \ldots & r_{3n} \\ & \vdots & & & \\ & & \vdots & & \vdots \\ r_{kn} & r_{2n} & r_{3n} & \ldots & 1 \end{vmatrix}}$$

which again enables the computation of all the R's of orders less than "n" as well.*

Similarly, this process allows the calculation of the standard errors of estimate of all orders up to and including "n," as given by the formula

$$\sigma^2_{1.23\ldots n} = \sigma_1\sqrt{1 - R^2_{1(23\ldots n)}}$$

---

[*] The row and the column where r_{11} appears must be stricken out in both determinants. The same applies to equation (2).

THE EFFECT OF THE 6 22 44 22 6 NORMAL CURVE SYSTEM ON FAILURES AND GRADE VALUES

J. DE WITT DAVIS

University of Oregon

There appeared recently a study[1] setting up the 6 22 44-22-6 per cent procedure for a five point grading scheme It has distinct theoretic advantage over those recommended by others, among them Dearborn[2] and Rugg[3] Any use of the normal curve as a basis for grade distribution arouses both strong opposition particularly from those who do not know how to apply it, and equally strong support from those who have given it the most study and most careful application. The growing employment of one or another modification of it prompts this analytical study.

The purpose here is not to show how either the small or large class as such is affected. The task assumed is twofold, namely:

1. Aside from other factors of test validity, and reliability, and also aside from other causes for leaving school, and aside from entrances other than as a freshman, what is the effect of a strict application of the 6-22-44-22-6 system as a basis for college grades from the standpoint of college failure?

2. And further, what is the comparative value or relative worth of grades from class to class when this system is applied?

For convenient thinking let us assume that after the weeding out of Freshman Week there remain a group of one thousand students for college work in a course which continues consecutively throughout the four college years with three terms in each year. No others join this purely hypothetical group, none leave it except by graduation as seniors, or because of being failed by the system. How many survive at the various levels?

The adopted system requires a brief explanation. By nature it forces each new term group into a normal distribution. It automatically fails six per cent, gives twenty-two per cent a grade of IV, forty-four per cent a grade of III, twenty-two per cent a grade of II and six per cent a grade of I. Since the group is forced into normalcy it may be assumed that practically one hundred per cent is included within $\pm 3\sigma$ range. AI then begins at approximately 1.6σ and takes the balance up; AII begins at a $+.6\sigma$ and extends up to the former; AIII begins at a $-.6\sigma$ and goes up to the lowest II; AIV begins at a -1.6σ and extends to the lowest III; while AV takes the balance, -1.6σ down.

---

[1] Eells, Walter Crosby: An Improvement in the Theoretical Basis of the Five Point Grading System Based on the Normal Probability Curve *Journal of Educational Psychology*, Vol. XXI, No. 2, Feb., 1930, pp. 128 130

[2] Dearborn, W. F.: School and University Grades *University of Wisconsin Bulletin*, No 368, Madison, Wisconsin, 1910.

[3] Rugg, Harold O.: "Statistical Methods Applied to Education " Houghton Mifflin Co., Boston, 1917, pp 216-219

A table constructed on the basis of this system shows clearly the elimination that takes place each of the twelve school terms when there are three terms per year for four years.

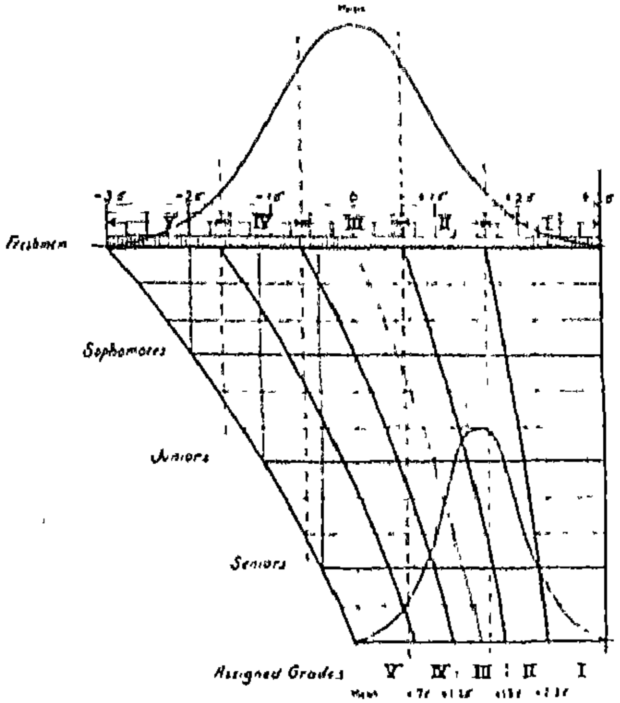
Table of Failures per School Term

Term	Entering	Number failing*	Accumulated failures
Freshman			
1st	1000	60	60
2nd	940	56	116
3rd	884	52	168
Sophomore			
1st	832	49	217
2nd	783	46	263
3rd	737	44	307
Junior			
1st	693	41	348
2nd	652	39	387
3rd	613	36	423
Senior			
1st	577	34	457
2nd	543	32	489
3rd	511	30	519
Total graduates	481	519	519

* Fractions are interpreted in favor of the students, *e g* 29½ = 30 students.

Assuming that the measures employed are perfectly valid and perfectly reliable, the poorer half have all been eliminated. This may or may not be desirable, a Platonic question might easily arise; if desirable, why not eliminate them at first and save the waste in time and expense both to the student and to the state? The answer might then come back that the measures really aren't perfect, besides, it is held by some to be essential that a student try out for a special ability in some given lines These are only implications aroused by the practice and therefore are only parenthetical to the question.

The table above answers the first query, namely, a strict application of the 6-22-44-22-6 plan gradually eliminates by failing a total fifty-one per cent of the admitted freshmen, sixty at the first term level and thirty at the last term of senior work.



6-22-44-22-6 Normal Curve Application

As to grades, what is their relative value from term to term, when assigned in this manner? By the use the geometric progression formula[1] the nomograph following was constructed Since the upper end of the distribution was not affected by failure the left or lower end was shortened. The border lines of the grade spreads as given above were determined by the use of Holzinger's Tables.[2] Notice on the nomograph that the mean at each level, by conforming to the normal distribution is caused to migrate to the right during the twelve terms a total distance of 1.5σ on the original base line. It is from this new mean always that the fixed distribution is made each term. A brief study of the graph makes the results evident

---

[1] $t_n = ar^{n-1}$ when t = the term wanted, n the number of the term in the series, a the first term of the series, and r the ratio.

[2] Holzinger, Karl J.: "Statistical Tables for Students in Education and Psychology" The University of Chicago Press, 1925.

A table of compared grades for the four classes as pictured by this graph will be sufficient.

FIRST TERM OF EACH YEAR
Comparison of Grade Values

Freshmen grades	Sophomores equivalents	Juniors equivalents	Seniors equivalents
I	All of Is Upper IIs	All of Is Most of IIs	All of Is All of IIs A few IIIs
II	Lower IIs About one-half of IIIs	Lower of IIs Nearly all of IIIs	Balance of IIIs Over one-half of IVs
III	Balance of IIIs Most of IVs	Balance of IIIs All of IVs Over one-half of Vs	Balance of IVs All of Vs
IV	Balance of IVs Most of Vs	Balance of Vs	
V	Balance of Vs		

Similar tables of comparison can be worked out by the interested reader using the grade of any other level as a criterion. It is evident that the farther removed from each other by school terms the grades may be the less comparable they become.

CONCLUSION

1. A strict application of the 6-22-44-22-6 normal curve procedure as a basis for grades in a twelve term course would eliminate practically fifty-one per cent of the matriculated group.

2. This strict practice changes the values of grades from term to term so that a grade of I or II, etc. has little directly comparable meaning unless associated with a particular term level

Much has been written about predicting grades from those made in former work, this study seems to indicate, though the measures were assumed to be perfectly valid and reliable, that predictability would yet be low due to this lack of comparability especially so when term grades of seniors are predicted say from freshman or sophomore grades. Readers may desire another system to provide for apparent skewness, at the same time to employ the basic technique of the normal curve and yet to keep the value of grades equal but that is another study. Others will recall this-and-that college which still has the two semester year division and will at once start comparisons with the twelve term schools. Many such implications arise calling for further consideration.

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XXII December, 1931 Number 9

THE ROUTINE COMPUTATION OF PARTIAL AND MULTIPLE CORRELATION

RAYMOND FRANZEN AND MAHEW DERRYBERRY

American Child Health Association, New York City

Probably the simplest and most direct method of computing multiple correlation and the regression coefficients of a multiple regression equation is the formation of a system of normal equations from the moments and product moments and the solution of these equations by the *Doolittle method as reported by Tolley and Ezekiel* in Vol. XVIII of the *Journal of the American Statistical Association.* A modified form of this method to be used when the intercorrelations are known and the raw moments are not available was published by Garrett.* It often happens, however, that not only are the regression coefficients *and the multiple correlation desired*, but the partial correlations are needed for interpretative purposes. In such problems it is more advantageous to compute the multiple correlation from the partial correlations according to the usual formula:

$$R_{1.23\ldots n} = \sqrt{1-(1-r^2_{12})(1-r^2_{13.2})(1-r^2_{14.23})\cdots(1-r^2_{1n.23\ldots n-1})}$$

and the regression coefficients from the formula:

$$b_{12.34\ldots n} = r_{12.34\ldots n}\frac{\sigma_1\sqrt{(1-r^2_{12})(1-r^2_{13.2})(1-r^2_{14.23})\cdots(1-r^2_{1n.23\ldots n-1})}}{\sigma_2\sqrt{(1-r^2_{12})(1-r^2_{23.1})(1-r^2_{24.13})\cdots(1-r^2_{2n.13\ldots n-1})}}$$

But the computation of $r_{1n.234\ldots n-1}$ and the other higher order partials becomes increasingly laborious as the number of secondary subscripts are increased. It can be shown that the formula

$$n + 2(n-1) + 3(n-2) + \cdots n(n-\overline{n-1}) \qquad (1)$$

---

* A Modification of Tolley and Ezekiel's Method of Handling Multiple Correlation Problems. *Journal of Educational Psychology*, January, 1928.

gives the minimum number of partials that are necessary for computing a partial of the *n*th order, or, in other words, it gives the minimum number of times the usual formula

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r^2_{13}}\sqrt{1 - r^2_{23}}}$$

must be applied to finally obtain a partial of the *n*th order. It will be noted that formula (1) gives the *minimum* number of computations. The possible number that may be computed is many more than that given by formula (1). For example, to compute a fourth order partial it is possible to compute thirty-five separate partials before finally obtaining a fourth order partial. On the other hand, only twenty partials need be computed if the most economical combinations of correlations are made in the computation of the lower order coefficients. Even though the computations are made in such a way that the minimum number of partials are computed, the usual method still requires considerable attention to avoid an incorrect combination in one of the lower order partials which would make all the subsequent work erroneous.

The method here presented completely eliminates these difficulties by reducing the solution to mechanical routine. It claims the following advantages:

1. The *n*th order partial correlation is obtained in the minimum number of operations, since the possibility of duplication of computations is eliminated.

2. Correct combinations of lower order partials are routinely made assuring the accurate higher order partial.

3. All computations are automatically checked without a double computation as is often advised.

4. Repeated writing of figures is eliminated.

5. Because of the direct mechanical and systematic calculation all the computations from first order correlations to the multiple can be done by a clerk.

6. A compact, concise record of the computations is provided for filing so that all partials of a given order and denomination are kept on the same page.

7. The multiple correlation is obtained from the partial forms with a minimum amount of computation.

The explanation below may seem difficult and complicated, but the routine is simple and easily followed once the method is under-

DIAGRAM I

(1)	r_{12}			
(2)	r_{15} r_{25}			
(3)	$r_{12} - r_{15}r_{25}$			
(4)	k_{15} k_{25}			
(5)	$r_{12.5}$			
(6)	r_{13}	r_{23}		
(7)	r_{15} r_{35}	r_{25} r_{35}		
(8)	$r_{13} - r_{15}r_{35}$	$r_{23} - r_{25}r_{35}$		
(9)	k_{15} k_{35}	k_{25} k_{35}		
(10)	$r_{13.5}$	$r_{23.5}$		
(11)	r_{14}	r_{24}	r_{34}	
(12)	r_{15} r_{45}	r_{25} r_{45}	r_{35} r_{45}	
(13)	$r_{14} - r_{15}r_{45}$	$r_{24} - r_{25}r_{45}$	$r_{34} - r_{35}r_{45}$	
(14)	k_{15} k_{45}	k_{25} k_{45}	k_{35} k_{45}	
(15)	$r_{14.5}$	$r_{24.5}$	$r_{34.5}$	
(16)	r_{15}	r_{25}	r_{35}	r_{45}
(17)	k_{15}	k_{25}	k_{35}	k_{45}
(18)		(Check 1)	(Check 2)	(Check 3)
(19)		$r_{15} + r_{25} + r_{35}$ (Check 4)	$r_{45} + r_{35}$ (Check 5)	r_{45} (Check 6)
(20)		$k_{15} + k_{25} + k_{35}$	$k_{35} + k_{45}$	k_{45}
(21)	$r_{12} + r_{13} + r_{14}$	$r_{23} + r_{24}$	r_{34}	
(22)	(r_{15}) (Check 1)	(r_{25}) (Check 2)	(r_{35}) (Check 3)	
(23)	Σ Numerator	Σ Numerator	Σ Numerator	
(24)	(k_{15}) (Check 4)	(k_{25}) (Check 5)	(k_{35}) (Check 6)	

DIAGRAM II

$r_{12.5}$		
$r_{13.5}$	$r_{23.5}$	
$r_{14.5}$	$r_{24.5}$	$r_{34.5}$

stood. The computational form with the entries and checks in terms of symbols is given in Diagram I. Here the operations necessary to compute a set of first order partials from the interrelationship of five variables are indicated. Diagram II denotes the method of transferring these first order partials to a second form for the computation of second order partials.

To make the diagrams more concrete the multiple correlation of four bony measurements with weight and two regression coefficients have been computed from the table of intercorrelations given below, and each step in the computation explained. The use of Diagrams I and II with the explanation will reduce the seeming complexity of the detailed directions that follow.

Table I.—Intercorrelations of Anthropometric Traits in One Thousand Eleven-year-old Girls*

	Weight	Height	Chest breadth	Chest depth
Height	7428			
Chest breadth	8437	6298		
Chest depth	8201	6488	6845	
Hip width	6752	7344	7824	6050

* Data gathered in connection with the School Health Study conducted by the American Child Health Association where this routine was devised and used.

Steps in the Computation

1. To compute first order partials the *r*'s are entered on specially ruled sheets (see Diagram III) just as they would be entered in an inter-*r* table with the exception that four lines are left below each *r* for the entry of computational figures. The order in which the coefficients are entered determines the order in which the variables will be partialled out, and before beginning computation this order must be decided upon so that the *r*'s will be so arranged on the sheet as to take care of this demand. The correlations of all other traits with the variable to be partialled out first are entered as the last row of intercorrelations, the correlations of all other traits with the variable to be partialled out second is placed second from the bottom, etc. Thus in this computation we have chosen to partial out hip width from all the remaining inter-*r*'s first and chest breadth second. (Compare Table I and Diagram III noticing arrangement of variables.)

2. Immediately below each of the correlations in the last line (16) of intercorrelations (the correlations with the variable to be partialled out first) is written the k or $\sqrt{1 - r^2}$ of the coefficient appearing directly above it (line 17). These values of $\sqrt{1 - r^2}$ are taken from "Tables of $\sqrt{1 - r^2}$ and $1 - r^2$" by Miner.

DIAGRAM III*

		I Weight	II Height	III Chest Depth	IV Chest Breadth
Height	(1)	7428			
	(2)				
	(3)				
	(4)				
	(5)				
Chest Depth	(6)	8201	5488		
	(7)				
	(8)				
	(9)				
	(10)				
Chest Breadth	(11)	8337	6208	6845	
	(12)				
	(13)				
	(14)				
	(15)				
Hip Width	(16)	8752	7334	6950	7323
k	(17)	483761	.079707	718420	680982
	(18)				
	(19)		2		7323
	(20)		2		680982
	(21)	2 3066	1 1080	6845	
	(22)				
	(23)				

* To simplify the explanation in the text the lines and columns are numbered. The numbers are not necessary in the ordinary computational procedure.

The computations described in Steps 3, 4 and 5 below are used in the check and are not necessary for the computation of the partials.

3. Beginning at the extreme right of the last row of intercorrelations (line 16) add each figure towards the left recording each cumulative total three lines below (line 19), and in the column corresponding to the last figure used in the cumulative total. Thus, 7323 is written

in the first column .7323 + .6956 is written in the second column, and 1.4279 + .7334 is written in the third, and so on towards the left until all of the r's have been added except the r in the last column to the left. It is not necessary to include this r in the summation.

4. The next line (line 20) is obtained from the line of k's (line 17) beginning similarly with the figure at the extreme right, and adding each figure in succession with the exception of the k in the last column to the left.

5. Sum each column of inter-r's omitting the last line of r's (line 16) and record on line 21. For example, 2.3966 is the sum of .7428 + .8201 + .8337.

Computation steps 6 to 11 are illustrated on Diagram IV.

6. If a calculating machine is used put the r that is in column I and in the bottom row (.8752 in row 16) in the machine and multiply by each succeeding r in that row. The products are written beneath each of the zero order r's in column I in succession. Thus the .6419 entered on line 2 is .8752 × .7334, .6088 entered on line 7 under the next zero r in the first column is the product of .8752 × .6956 and .6409 on line 12 is .8752 × .7323.

7. Similarly put the r in column II and in the bottom row (.7334 in row 16) in the machine and multiply by each succeeding r in that row and record the products in column II in succession on lines 7 and 12. The entry in column III, line 12, is the product of the r in column III on the last line (.6956) and the r in column IV (.7323). These computations give the encircled term of the formula below for each of the six partials on Diagram IV.

$$r_{xy\cdot z} = \frac{r_{xy} - r_{xz}r_{yz}}{k_{xz}k_{yz}}$$

8. Subtract each of these products obtained in Step 7 from the r just above it in the column. Thus, .1009 in line 3 is .7428 − .6419. (The operations indicated are, of course, always performed algebraically, giving proper regard for signs.) This is the numerator of the formula given above.

To check result thus far:

9. Multiply the r in the bottom row of column I (line 16) by the figure in column II, line 19, and enter the product in column I, line 22. Thus, 1.8916 is the product of .8752 × 2.1613. Likewise multiply the r in the last row of column II by the figure in column III, line 19, and record the product in column II, line 22; i.e., the 1.0173 is the

product of .7331 × 1.1279. The remaining product in column III, line 22, is found in a similar manner.

10. Subtract the products obtained in Step 9 from the summations immediately above them on line 21, and enter the differences in line 23.

11. Sum the numerators in column I on lines 3, 8 and 13 obtained in Step 8. The sum should equal the figure shown in column I, line 23, as the result of Step 10, (i.e. .1009 + .2113 + .1928 = .5050).* Of course in some cases there may be a difference of 1 or 2 points in the last decimal place due to dropping of decimals. In a similar manner columns II and III are checked.

To compute the denominators:

12. Perform the multiplications described in Step 6 using the row of k's, line 17, instead of r's, line 16, and enter the products on the lines below the several numerators obtained in Step 8. This gives the denominator of the formula above, (i.e. .3289 = .183761 × .679797).

To check these entries:

13. Multiply the k in column I by the summation in column II on line 20. Record in column I, line 24. Sum the denominators in column I obtained from Step 12, lines 4, 9 and 14. The two results should check. Similarly multiply the k in column II by the summation in column III and the product should equal the sum of the products of k's in column II.

Final computations:

14. Divide the numerators obtained in Step 8 by the respective denominators obtained in Step 12. The results are usually written in red (italics here) or otherwise made distinctive so they can be picked out readily. This quotient is the partial. For example, .1009 ÷ .3289 gives .3068, which is the partial correlation of height and weight independent of the influence of width of hips. Similarly .3579 is the relation of chest breadth to chest depth independent of hip width. The only check on this division is the multiplication of the denominator by the quotient, obtaining the numerator.

To compute second order partials.

15. The first order partials obtained on Diagram IV are entered as shown in Diagram V. The arrangement is the same as Diagram III except that no entries are made on line 16, since all the correlations

---

* This may be shown to be true by the following equation

$$(r_{12} - r_{15}r_{25}) + (r_{13} - r_{15}r_{35}) + (r_{14} - r_{15}r_{45}) = r_{12} + r_{13} + r_{14} - r_{15}(r_{25} + r_{35} + r_{45})$$

with hip width have been partialled out. The last line is now chest breadth. These entries are checked by summations of the successive columns of Diagram IV agreeing with the summations of the corresponding column of Diagram V. By performing the computations described in Steps 2 to 11, the variable chest breadth will be partialled

Diagram IV*

		I Weight	II Height	III Chest Depth	IV Chest Breadth
Height	(1)	7428			
	(2)	6419			
	(3)	1009			
	(4)	3289			
	(5)	*8068*			
Chest *D*	(6)	8201	5488		
	(7)	6088	5102		
	(8)	2113	0286		
	(9)	3175	4884		
	(10)	*6081*	*6586*		
Chest *B*	(11)	8347	6298	6845	
	(12)	6409	5371	5094	
	(13)	1028	0027	1751	
	(14)	3204	4629	4892	
	(15)	*5863*	*8008*	*3579*	
Hip Width	(16)	8752	7341	6950	7323
k	(17)	483761	679707	718429	680982
	(18)				
	(19)		2 1613	1 4270	*7323*
	(20)		2 079208	1 399111	680982
	(21)	2 3066	1 1686	6845	
	(22)	1 8916	1 0173	5094	
	(23)	5050	1213	1751	
	(24)	1 0058	9513	4892	

* Diagram IV is like Diagram III with the remaining computations completed

out. The results obtained from these computations are all relationships of two of the variables irrespective of hip width and chest breadth.

Each of the successive higher order partials can be obtained by repeating the process, partialling out one additional variable with each repetition.

DIAGRAM V

	I Weight	II Height	III Chest Depth
(1)	3008		
(2)	1172		
(3)	1806		
(4)	7044		
(5)	2387		
(6)	6081	0580	
(7)	2091	0717	
(8)	3986	− 0131	
(9)	7571	0148	
(10)	*5966*	− *0143*	
(11)	5953	2004	3570
(12)	810817	070735	033760
(13)			
(14)		5582	3570
(15)		1 913405	033760
(16)	9140	0580	
(17)	3267	0717	
(18)	5882	− 0131	
(19)	1 5515	0148	

Computation of 3rd Order Partials

(a) I Weight			(b) I Weight	
2387		Chest *D*	5205	
− 0075			− 0034	
2462			5299	
8501			9710	
2800	II Height		*5457*	II Height
5205	0143	Height	2387	−.0143
850175	000808		071003	000808

Computation of the second and third order partials are pictured in Diagram V. The third order partial $r_{12\cdot 345}$ is obtained in the usual manner. The third order partial $r_{13\cdot 245}$ is also given in this diagram and will be used later in the computation of the regression coefficients. It is obtained from a rearrangement of the second order correlations. This is always possible when computing partials of the order n-2 (where n is the number of variables involved), for

$$r_{12\cdot 34\ldots n} = \frac{r_{12\cdot 45\ldots n} - (r_{13\cdot 45\ldots n})(r_{23\cdot 45\ldots n})}{(k_{13\cdot 45\ldots n})(k_{23\cdot 45\ldots n})} \text{ and}$$

$$r_{13\cdot 24\ldots n} = \frac{r_{13\cdot 45\ldots n} - (r_{12\cdot 45\ldots n})(r_{23\cdot 45\ldots n})}{(k_{12\cdot 45\ldots n})(k_{23\cdot 45\ldots n})}$$

To compute the multiple correlation.

The multiple correlation coefficient may be expressed as a function of the alienation coefficients of the partial correlation, thus:

$$R_{1\cdot 234\ldots n} = \sqrt{1 - (k_{12})(k_{13\cdot 2})(k_{14\cdot 23}) \ldots (k_{1n\cdot 23\ldots n-1})}$$

If the criterion variable has been entered in column I in the computation of the successive higher order partials all of the k's required in the above formula have been determined except $k_{1n\cdot 23\ldots n-1}$. But $r_{1n\cdot 23\ldots n-1}$ has been computed so the corresponding k can be quickly determined from Miner's Tables. In the illustrative table the k of .2896 is .957148. By a multiplication of the k's in the first column on each successive sheet, and the k of the highest order partial in column I the multiple k or ($k_{1\cdot 234\ldots n}$) will be obtained. Thus $k_{1\cdot 2345}$ = .483761 (Diagram IV, line 17) × .840847 (Diagram V, line 12) × .850175 (Diagram V, line 7) × .957148 (k of .2896) = .3192. From this product $R_{1\cdot 2345}$ which, of course, is $\sqrt{1 - k^2_{1\cdot 2345}}$ is obtained from Miner's Tables. In this example $R_{1\cdot 2345}$ = .9477.

To compute the regression coefficient.

If we substitute in the multiple regression equation

$$\bar{X}_1 = b_{12\cdot 345}X_2 + b_{13\cdot 245}X_3 + b_{14\cdot 235}X_4 + b_{15\cdot 234}X_5 - C \qquad \text{(A)}$$

for $b_{12\cdot 345}$, $b_{13\cdot 245}$, etc. the partial r's and the partial σ's that are their equivalents the equation reduces to

$$\begin{aligned} \bar{X}_1 = {} & r_{12\cdot 345}\frac{\sigma_1}{\sigma_2}\frac{k_{1\cdot 2345}}{(k_{25})(k_{24\cdot 5})(k_{23\cdot 45})(k_{12\cdot 345})}X_2 \\ & + r_{13\cdot 245}\frac{\sigma_1}{\sigma_3}\frac{k_{1\cdot 2345}}{(k_{35})(k_{34\cdot 5})(k_{23\cdot 45})(k_{13\cdot 245})}X_3 \\ & + r_{14\cdot 235}\frac{\sigma_1}{\sigma_4}\frac{k_{1\cdot 2345}}{(k_{24})(k_{34\cdot 2})(k_{45\cdot 23})(k_{14\cdot 235})}X_4 \\ & + r_{15\cdot 234}\frac{\sigma_1}{\sigma_5}\frac{k_{1\cdot 2345}}{(k_{25})(k_{35\cdot 2})(k_{45\cdot 23})(k_{15\cdot 234})}X_5 - C \qquad \text{(B)} \end{aligned}$$

If, in terms of the illustrative problem, we let X_1 = weight, X_2 = height, X_3 = chest depth, X_4 = chest breadth and X_5 = hip width, then it can readily be seen that all the partial r's and partial k's in the first two terms of equation (B) are taken care of by the computations on Diagram IV and Diagram V From Diagram IV we get k_{15}, k_{25}, k_{35}, and from Diagram V we get $k_{24.5}$, $k_{34.5}$, $k_{23.45}$, $r_{12.345}$ and $r_{13.245}$

In order to compute the partials necessary for the two remaining terms it is necessary to make the same number of computations as those made in Diagrams IV and V, *reversing the order in which the* variables are partialled out, in terms of the illustrative problem, partial out the variables in the order height, chest depth and chest breadth in addition to the order hip width, chest breadth and chest depth as illustrated on the diagrams presented Such computations would yield the partials required for the remaining two terms of equation (B).

In problems involving more than 5 variables it is only necessary to go from zero order r's to the $(n - 2)$ order partials one time in order to obtain the Multiple R. If the regression equation is desired it is necessary to compute four $(n - 2)$ order partials in the manner indicated above and then return to the second order r's, rearrange them on the work sheet and then compute other partials of order $(n - 2)$. For example, if the problem involves 7 variables, and in the first computation the variables are partialled out in the order 7, 6, 5, 4, 3, giving the two partials $r_{12.34567}$ and $r_{13.24567}$, and if in the second computation the variables are partialled out in the order 2, 3, 4, 5, 6, giving the partials $r_{17.23456}$ and $r_{16.23457}$, the remaining two partials desired could be obtained by partialling out in the order 2, 3, 7, 6, 4, giving $r_{14.23567}$ and $r_{15.23467}$ But the computations necessary to partialling 2 and 3 out were done in the second computation, so a rearrangement of the sequence on the computational form of the second order r's is all that is necessary for these additional higher order r's This again reduces the amount of computation necessary.

Though the explanation may seem long and complex, the routine is very simple. After it is understood and has become mechanical in operation, the time saving is approximately one-half of the usual method. It can be explained to clerks verbally much more easily than has been possible in this written explanation

A STUDY OF QUESTIONNAIRE TECHNIQUE

FRANK K. SHUTTLEWORTH

Yale University

This study shows that under the conditions defined below the enclosure of a twenty-five cent piece in a mail questionnaire brought 32.4 per cent more replies than the same questionnaire without the coin. Six hundred eight persons receiving the coin returned 51.6 per cent replies while three hundred seventy-six not receiving the coin returned only 19.1 per cent Data are also presented on the rôle of selective factors and on relative costs.

The study is a by-product of an evaluation of the Cattaraugus County Health Demonstration in western New York. For a period of seven years the Milbank Memorial Fund financed an intensive demonstration of rural public health work in Cattaraugus County. In the spring of 1930 a staff of experts under the direction of Professor C-E. A. Winslow[1] undertook to evaluate that work. The author assisted in a survey of the educational aspects of the demonstration among school children and adults. With some reluctance a questionnaire by mail was used to assay the attitudes of adults especially toward the financial support of public health work. One of our chief doubts about the questionnaire was the fear of an inadequate proportion of replies. To solve this difficulty, it was decided to enclose a twenty-five cent piece secured in a coin mailing card with each questionnaire. Evaluation of the results required a control county and for this purpose, Steuben, the county just east but one from Cattaraugus, was selected. The inquiry to these counties consisted of a short personal letter each of which was separately typed and obviously so, the questionnaire proper containing six questions, a stamped and addressed envelope, and the coin.

The investment involved and the novelty of the procedure naturally led to an attempt to test the effect of the coin. Accordingly, the inquiry was duplicated in all respects save only two sentences in the personal letter and the presence of the coin. This inquiry was sent to persons in the six New York counties immediately surrounding Cattaraugus and Steuben.

The personal letter which accompanied the coin was as follows:

---

[1] Winslow, C-E A., "Health on the Farm and in the Village" Macmillan, 1931.

> A responsible organization has asked me as an impartial outsider to make a survey of opinion toward public health work in the western counties of New York State. They are anxious to have 100 per cent replies to the questions and accordingly have provided funds which make possible the enclosed coin. It is not in payment for your trouble in answering the questions but rather a small token of appreciation of your coöperation.

The letter not accompanied by a coin substituted the following for the last two sentences in the above letter:

> The results of such a survey will be trusted to the extent that every one coöperates in answering the enclosed questions. I will be grateful if you will indicate your opinion and send it to me quite promptly in the enclosed stamped and addressed envelope.

These letters were written on the stationery of the Department of Education of Yale University.

The comparability of the two areas to which these inquiries were sent is of importance in evaluating the influence of the coin. In connection with the health study, which was searching for differences between Cattaraugus and Steuben, all available population, economic, and educational statistics were examined. On thirty-seven basic factors the two counties proved to be very much alike. Cattaraugus is, of course, very different from Steuben in one respect having had a well organized and adequately financed county health program for seven years while Steuben had had no organized county health program. However, the replies to the questionnaire showed only slender advantages on the part of the Cattaraugus respondents. For the purpose of this study we feel justified in combining them into one area which will be referred to as the coin area. Additional data testing the comparability of the coin area and the non-coin area was not examined for the reason that only the portions of the six counties most immediately adjacent to Cattaraugus and Steuben were canvassed. Instead, reliance is placed on the similarity between Cattaraugus and Steuben which are separated by the intervening county of Allegheny. The complete details are recorded in the accompanying table in the order in which they are discussed in the text. Data for the coin area are also given separately for Cattaraugus and Steuben counties.

The inquiry itself provided other controls. The questionnaires were sent only to names listed in village telephone directories, thus introducing an economic control, and approximations to social and cultural controls. In the non-coin area sixteen village exchanges serving forty-two still smaller hamlets and cross-roads were sampled. In the coin area thirty village exchanges serving fifty-seven additional hamlets were sampled. The sixteen villages in the non-coin area average 8.6 miles from the borders of the coin area. All of the questionnaires were mailed on the same day.

Questionnaires were mailed to every eighth name on these exchanges excepting the names of business firms and of women. This gave three hundred eighty letters to the non-coin and six hundred seventeen to the coin area. Four and nine letters were returned for insufficient address. Of the three hundred seventy-six and six hundred eight persons who presumably received the inquiry, seventy-two or 19.1 per cent from the non-coin area and three hundred fourteen or 51.6 per cent from the coin area made some reply. The difference of 32.4 per cent is sixteen times as large as its probable error.

The coin area returned 2.7 times as many replies. This ratio gradually increased throughout the period during which replies were coming in. The first three days with a total of one hundred fifty-three replies showed a ratio of 2.5, the second three days with one hundred fifteen replies showed a ratio of 2.6, while the second, third, and fourth to tenth weeks showed ratios of 2.7, 3.9, and 4.9.[1]

A detailed analysis of the responses is presented since it throws light on two problems of the questionnaire method. One of the haunting ghosts of this procedure is the selection which presumably is involved in the fact that some persons reply and others do not. Presumably less selection is involved in 52 per cent replies than in 19 per cent. On the other hand there is the question whether the presence of the coin tended to disturb the answers. The data offer many tests of these questions although in the nature of the case it is impossible to say whether one or both factors are responsible for any differences. Interpretation of similarities is equally difficult since they may mean that neither selection nor the coin disturbed the answers or that these two factors tended to cancel each other.

A preliminary point of some interest is that fourteen of the three hundred fourteen respondents from the coin area returned the coin. Seven of these filled out the questionnaire in detail and by comments or implication indicated that they thought the coin unnecessary while seven returned the coin without the questionnaire or with the questionnaire left blank. Although the questionnaire said specifically "You do not need to sign your name" twelve of the fourteen or 86 per cent gave their names while among the other three hundred respondents only fifty-two or 17 per cent gave their names.

---

[1] Subsequent to the preparation of this article ten weeks after mailing the questionnaires an additional ten replies were received up to the end of the fourth month. Nine of these were from the coin area and one from the non-coin area. These additional cases are not included in this analysis.

The sixth question in the inquiry asked whether the respondent's taxes on real estate amounted to over $100 the previous year. In the non-coin area, of sixty-eight answering this question, 63.2 per cent reported more than $100 taxes. In the coin area of two hundred eighty-five answering this question 59.5 per cent made this report. The difference is 3.7 ± 4.6 per cent. All of the data were tabulated separately for those reporting more and for those reporting less than $100 taxes but since no differences in the answers of the two groups appeared they are combined in the discussion which follows.

The first question in the inquiry asked the respondents to rank four functions of public health work in order of importance. Forty-eight from the non-coin area and one hundred fifty-eight from the coin area properly followed instructions in ranking the four functions. The percentages of correct rankings for function (a) were 47.9 and 44.3 per cent, for function (b) 81.3 and 77.9 per cent, for function (c) 27.1 and 36.7 per cent, for function (d) 51.2 and 51.9 per cent. The largest difference here is 9.6 ± 5.3 per cent.

The second, third, fourth, and fifth questions concerned financial support of public health work. Favorable attitudes were expressed by answering "No" to the second question and "Yes" to the third, fourth, and fifth. For the non-coin and coin areas the percentages of favorable replies were 79.4 and 77.1 per cent, 70.5 and 70.4 per cent, 41.9 and 39.5 per cent, and 73.8 and 77.2 per cent. Here the largest difference is only 3.4 per cent.

We turn next to the care with which the inquiry was answered by the two areas as indicated by the proportion of omitted or uncertain answers. Instead of seventy-two and three hundred fourteen cases as our base we shall use seventy-one and three hundred seven since one respondent from the non-coin area wrote that he had mislaid his questionnaire and seven from the coin area returned the coin without answering any questions. For the six questions the non-coin area averages 15.0 per cent omissions or uncertain answers while the coin area averages 17.0 per cent. Analysis by specific questions, however, reveals one significant difference. Only 32.9 per cent from the non-coin area failed to rank the four functions of public health as instructed in comparison with 50.4 per cent from the coin area. The difference of 17.5 per cent is 4.1 times as large as its probable error. On two other questions the coin area showed more omissions while on the other three it showed fewer omissions. The largest of these differences is only 6.2 per cent

The respondents from the two areas were equally ready to back their replies with their signatures. Although the questionnaire specifically said, "You do not need to sign your name," 15.5 per cent from the non-coin area signed their names. After deducting the fourteen cases who returned the coin and naturally wanted to receive credit for it there remain 18.0 per cent of the respondents from the coin area who gave their names

A final test of the influence of selection and of the coin is the interest indicated by comments. These were invited by "Use the reverse side for comments" following the sixth question From the non-coin area 44.5 per cent made comments while from the coin area only 33.2 per cent made any comments. The difference of 11.3 per cent is only 2.6 times as large as its probable error. While a smaller proportion of respondents from the coin area made comments, those who did so made more and longer comments. Counting each qualification to a specific question and each paragraph of discussion as a comment, the non-coin area averages 1.78 comments per person commenting and .79 comments per person replying. In the coin area these figures are 1.92 and .64. Counting all words in the comments gives an average of 30.40 per person commenting and of 13.50 per person replying in the non-coin area in comparison with 48.02 and 16.25 in the coin area. The most significant of these differences is only 2.2 times as large as its probable error. The differences are also exaggerated because these measures are the ones in which Cattaraugus approaches significant superiority over Steuben.

The relative costs of the two procedures are of some interest. No expense was spared to give the inquiry a favorable appearance The letters to the non-coin area cost 10.4 cents each while those to the coin area cost 36.5 cents. The initial expense to the coin area was 250 per cent greater per letter. The larger proportion of replies reduced this discrepancy very much. In costs per reply of any kind, the coin area exceeded the non-coin area by only 31 per cent In costs per relatively complete reply the excess for the coin area was only 32 per cent. In costs per reply containing one or more comments the excess was 80 per cent. In costs per comment the excess was 61 per cent

In costs per word of comment the coin area exceeded the non-coin area by only 12 per cent

SUMMARY

The enclosure of a twenty-five cent piece in a simple questionnaire sent out by mail brought 51.6 per cent replies while the same questionnaire without the coin returned only 19.1 per cent replies

TABLE I SUMMARY OF DATA

	Non-coin		Coin		Cattaragus		Steuben	
	No.	Per cent or *M*	No.	Per cent or *M*	No.	Per cent or *M*	No.	Per cent or *M*
Village telephone exchanges sampled	16		30		14		16	
Other hamlets included	42		33		21		12	
Inquiries mailed	380		617		287		330	
Returned for address	4		9		2		7	
Persons receiving inquiry	376		608		285		323	
Replies of any kind	72	19.1	314	51.6	151	53.0	163	50.5
Wrote that questionnaire was mislaid	1							
Returned coin with questions answered			7		6		1	
Returned coin without answering questions			7		4		3	
Useable returns	71	18.7	307	50.5	147	51.6	160	49.5
Reported less than $100 taxes	25	36.8	121	41.5	51	38.8	70	47.9
Reported more than $100 taxes	43	63.2	161	58.5	85	61.2	76	52.1
Correctly ranking function (a) fourth	23	47.9	70	44.6	39	47.6	31	41.3
Correctly ranking function (b) first or second	39	81.3	121	77.1	64	77.1	59	78.7
Correctly ranking function (c) third	13	27.1	58	36.9	29	34.0	29	38.7
Correctly ranking function (d) first or second	26	54.2	82	52.2	45	53.2	37	49.3
Correctly answering question No. 2 "No"	50	70.4	215	77.4	98	70.5	117	78.1
Correctly answering question No. 3 "Yes"	43	60.6	197	70.1	97	71.0	100	66.7
Correctly answering question No. 4 "Yes"	29	40.8	111	39.5	55	41.7	56	37.8
Correctly answering question No. 5 "Yes"	51	71.8	220	77.2	111	81.1	109	71.8
Uncertain or omitted answers to six questions	64	15.0	313	17.0	156	17.7	157	16.3
Uncertain or omitted tax report	3	4.3	29	9.5	12	8.2	17	10.6
Uncertain or omitted ranking of functions	23	32.4	156	50.8	68	46.3	88	55.0
Uncertain or omitted answers to question No. 2	8	11.4	36	11.7	23	15.6	13	8.1
Uncertain or omitted answers to question No. 3	10	14.1	34	11.1	21	14.3	13	8.1
Uncertain or omitted answers to question No. 4	9	12.8	33	10.7	18	12.2	15	9.4
Uncertain or omitted answers to question No. 5	10	14.1	25	8.1	11	7.5	14	8.8
Persons signing names	11	15.5	66		36		30	
Deducting fourteen cases returning coin, twelve of whom signed names			54	18.0	28	17.7	26	16.2
Persons making comments	32	45.1	102	33.2	53	36.1	49	30.6
Average number of comments per person commenting		1.78		1.92		2.21		1.61
Average *number of words per person commenting*		30.40		49.02		57.31		39.82
Average number of comments per person replying		.79		.64		.80		.49
Average number of words in comments per person replying		13.70		16.29		20.67		12.19
Total expense of sending inquiry in dollars		38.40		225.50		104.89		120.61
Costs per letter sent out in cents		10.1		36.5		36.5		36.5
Costs per reply of any kind in cents		53.3		71.8		69.5		74.0
Costs per relatively complete reply in cents		54.1		73.5		71.4		75.4
Costs per reply with any comment in cents		120.0		221.1		197.9		246.1
Costs per comment in cents		68.1		115.1		89.7		152.7
Costs per word of comment in cents		4.0		4.5		3.5		6.2

Detailed analysis of the returned questionnaires shows negligible difference between the non-coin and coin areas. On one question the respondents from the coin area showed significantly more omissions or uncertain answers, but considering omissions and uncertain answers to all questions the advantage of the non-coin area is only 2 per cent. A larger proportion of respondents from the non-coin area made comments, but the comments of respondents from the coin area tended to be longer.

From the absence of differences in the responses of the non-coin areas to the questionnaire, one of the following inferences is warranted. Neither the selection involved in a small proportion of replies nor the presence of the coin was disturbing to the replies. The influence of selection and of the coin cancelled each other.

While the initial cost of enclosing a twenty-five cent piece made the letters to the coin area 250 per cent more expensive than letters to the non-coin area, the larger returns from the coin area reduced the excess in terms of unit costs as low as 12 per cent. At the most the actual excess expense is 80 per cent per unit of returns.

The procedure of paying the victims of a questionnaire is recommended, not for adoption, but for serious consideration. In terms of costs it did not prove superior in this instance. Whether on the whole it was superior in the present case is a matter of judgment which must go beyond the actual data in evaluating the importance of selective factors and the possibility that the coin itself was a disturbing factor.

INFLUENCE OF THE ASSIGNMENT ON LEARNING

DAVID H. BRIGGS

Florida State College for Women

A. M. JORDAN

University of North Carolina

The more effective utilization of the assignment as an instrument of instruction has been the burden of much discussion but of little experimentation. Recent writers such as Odell, Crawford and McDonald, Stone, Sears, and Monroe have pointed out the need for adequate assignments of lessons. By "adequate" they imply that assignments should be clear, helpful as to study procedures, stimulating, and that they should arouse problems to be solved. With the outcomes of these procedures they have not concerned themselves. It has seemed to us worthwhile, therefore, to investigate the amount of added learning which accrues from certain assignment procedures.

The assignment procedures, five in number, were chosen because of their general use rather than because they are the best. These procedures are:

(a) The pupils were furnished a suggested study procedure which consisted of looking up meaning of difficult words, drawing lines under the important statements, and learning the significance of the statements previously underlined. Great care was taken to make sure that the pupils knew exactly what they were to do before they were allowed to begin.

(b) Carefully selected questions on the material to be studied were given to the pupils before they began studying. The instructions were "I want you to find the answers to these questions, and when you have found them draw a line under the correct answer. Try to learn as many other facts as you can while you are reading and studying the selection. You are to use these questions as a guide in helping you to master the points brought out in the selection."

(c) The pupils were taught the meaning of the difficult words in the material to be studied. A sheet was presented to the pupils containing a list of the more difficult words with their appropriate meanings with these instructions "After we have spent about ten minutes studying these hard words I am going to give you a test to see how many of them you have learned. Then, too, you will need to know their meanings in order to answer the questions you will be asked later on in the passage you are to read."

(d) Attempt was made to show the children the manner in which the tariff (the subject of the material studied) affected them personally. Illustrations were presented of types of goods made more cheaply in foreign lands than we could manufacture in this country because of the lower standards of living obtaining there than here. The ordinary high tariff argument was presented which emphasized the effect of lowering or raising the tariff on wages and family income and the consequent effect on the pupils themselves.

(*c*) In this case, there was a conscious attempt to explain what the tariff was, why some thought it a good measure, why it is called a protective tariff and why it was levied. It was emphasized that many of the pupils might differ from the President politically but that it was their duty to learn much about a topic discussed in Congress at that very time.

The material used was selected from the message of President Hoover on the general subject of the tariff. It consisted of two pages of single spaced typewritten material and was in nature more difficult than the ordinary reading matter used by pupils of the levels studied. The four pages were divided equally into two sections, *A* and *B*. This could be done rather easily and naturally since there was a break in the thought at the point of division. For each paragraph there were prepared three sorts of material. The first consisted of a series of words chosen from the more difficult words of the text with multiple choice answers. The second type of material was a test of sentence meaning. Care was exercised in making the answers depend on the understanding of the individual sentence and of that alone. The third type consisted of statements based on the interrelation and meaning of several sentences taken together. Sometimes these questions depended for their answer upon the interpretation of what was read. These three types are called: (1) Word knowledge, (2) sentence meaning, (3) paragraph meaning. Before the experiment proper, this material was checked for its difficulty (1) by counting the number of words in each selection more difficult than the first five thousand in "Thorndike's Word Book" and (2) by testing the material on two hundred forty-one cases located in Grades V, VII, and IX. This last procedure furnished us with the following results:

	Word knowledge	Sentence meaning	Paragraph meaning
Form A	13 11	6 06	4 47
Form B	14 03	5 54	4 10
Differences of means	92	52	32
PE$_{\text{diff}}$ of means	37	.14	12

The PE of the differences of the means showed the differences between *A* and *B* in sentence meaning and in paragraph meaning to be reliable. The difference was unreliable in the case of word knowledge

The differences between the means were provided for by presenting *A* and *B* in alternate manner. If there were two grades, in the one we used *B* for the control group and *A* for the experimental group; in the other, *A* was the control and *B*, the experimental. The correlation between the two forms was .80 ± .01

General Procedure

The five situations were tested in the Durham and Raleigh, N. C., schools. Two groups of fifth, two of seventh, and two of ninth grade pupils were subjects for each of the experimental situations. This made a total of one hundred sixty-seven subjects for Assignment *A*, one hundred eighty-six for Assignment *B*, two hundred fifteen for Assignment *C*, one hundred eighty-nine for Assignment *D*, and one hundred ninety-four for Assignment *E*: nine hundred fifty-one subjects in all. The same examiner (Briggs) gave all the assignments and all the tests to each of the thirty groups of pupils.

The general plan was to test the children first with one of the forms on one day, then the next day make the assignment with the other form and calculate the differences scored by pupils who on one day were influenced by certain types of assignments and on the other were not so influenced. The total time consumed by the pupils in these procedures was from thirty-five to fifty minutes

Results

The average results only are given on account of the lack of space

Table I Results for Assignment A
(Suggested Study Procedure)

	Word knowledge	Sentence meaning	Paragraph meaning
Control	15.58	6.34	4.73
Experimental	18.66	6.43	4.80
Number . . .	167	167	167
Difference of means	3.08	.09	.07
PE$_{\text{diff}}$ of means	.53	.23	.19

In studying Table I, one is struck by the slight differences between the means. In the sentence meaning and in paragraph meaning the differences are merely chance differences but in word knowledge there was a reliable gain due to the assignment. It is to be remembered that the very procedure of Assignment *A* almost forced an improvement in vocabulary since it consisted among other things of looking up the meaning of difficult words Since we had at our disposal the records of several grades it was not difficult to get an average gain for the year in each of the tests. The average gain per year on the vocabulary test was 3.07 words which is just about the same as the improvement of the vocabulary scores as a result of this procedure The mean gain per year for sentence meaning was .84 questions and for paragraph meaning .75 questions Although the meanings of the words were studied and learned sufficiently well to recognize their correct meaning among five words they were not learned sufficiently well in the time allowed for them to have entered into the warp and woof of the sentence in such a manner as to improve the scores on the sentences and paragraphs

TABLE II – RESULTS FOR ASSIGNMENT *B*
(Assignment of Pertinent Study Questions)

	Word knowledge	Sentence meaning	Paragraph meaning
Control	16.18	6.33	4.78
Experimental	16.33	6.53	4.88
Number	186	186	186
Difference of means	.15	.20	.10
PE_{diff} of means	.47	.21	.17

TABLE III – RESULTS OF ASSIGNMENT *C*
(Teaching Meaning of Difficult Words)

	Word knowledge	Sentence meaning	Paragraph meaning
Control	15.06	6.34	4.74
Experimental	26.09	6.60	5.03
Number	215	215	215
Difference of means	11.03	.26	.29
PE_{diff} of means	.44	.20	.16

It is clear from Table II that the assignment of pertinent study questions had little if any demonstrable result on word knowledge or paragraph meaning and a slight but unreliable effect on understanding the meaning of sentences.

It is apparent that the words learned in the experimental group did influence the scores in word knowledge. More interesting perhaps is the small improvement in sentence and paragraph meaning when so many of the words had been learned. The improvement looms larger, however, when compared with the yearly growth of .81 questions in the sentence meaning and of .75 in paragraph meaning. The difference between the experimental and the control in word knowledge needs no comment since these words were taught the children directly. In sentence meaning the chances are sixty-two in a hundred that if the experiment were repeated the experimental group would still be ahead of the control group. In the case of paragraph meaning similar chances are seventy-eight in a hundred. Teaching the meaning of difficult words, as far as our study goes, has a decided effect upon the ability to comprehend the meaning of sentences and paragraphs.

TABLE IV. RESULTS OF ASSIGNMENT *D*
(Making Pupils Aware of Personal Value of Materials)

	Word knowledge	Sentence meaning	Paragraph meaning
Control	17.87	6.96	5.43
Experimental	18.57	6.92	5.48
Number	189	189	189
Difference of means	.70	-.04	.05
PE_{diff} of means	.49	.22	.18

TABLE V. RESULTS OF ASSIGNMENT *E*
(Making Pupils Aware of the General Background of the Material)

	Word knowledge	Sentence meaning	Paragraph meaning
Control	17.32	6.77	5.48
Experimental	18.09	7.07	5.59
Number	194	194	194
Difference of means	.77	.30	.11
PE_{diff} of means	.17	.20	.16

As a result of making pupils aware of the personal value of the tariff, there was a fairly significant improvement in word knowledge. The effects on sentence meaning and paragraph meaning were nil

Making children aware of the general background of the material studied produces a positive effect in the three types of answer required of them. The differences are not statistically reliable but all of them are positive and in the same direction. Compare these differences with a year's growth of 3.08 in word knowledge, .81 in sentence meaning and .75 in paragraph meaning and it is clear that a few minutes of explanation brings about ¼ of a year's growth in word knowledge, ⅓ of a year's growth in sentence meaning, and 1/7 of a year's growth in paragraph meaning. It will be seen that the improvement is greatest in sentence meaning and least in paragraph meaning.

TABLE VI.—MEAN GAINS IN EACH OF THE SITUATIONS WITH THEIR PROBABLE ERRORS

	Word knowledge	Sentence meaning	Paragraph meaning	Number
(*a*) Suggested study procedure	3.08 ± .53	.00 ± .23	.07 ± .19	167
(*b*) Pertinent study questions	−.15 ± .17	.20 ± .21	.10 ± .17	186
(*c*) Teaching meaning of words	1.03 ± .41	.26 ± .20	.26 ± .16	215
(*d*) Making pupils aware of personal value of materials	.70 ± .40	−.01 ± .22	.05 ± .18	189
(*e*) Developing background	.77 ± .47	.30 ± .20	.11 ± .16	191
Average yearly gain	3.08	.81	.77	

Table VI summarizes the results of this investigation. It will be seen that the scores encircled are as large or larger than their corresponding probable errors. From the strictest statistical point of view only Assignments *A* and *C* are reliable for in these cases the differences are more than four times the probable error of the difference. But in these cases the teaching directly influenced the scores on Word knowledge. It is clear that the teaching of the meaning of words produced an improvement in sentence meaning of .26 of a question or a little less than ⅓ of a year's growth and slightly more than ⅓ of a year's growth in paragraph meaning. This careful teaching of the meaning of the more difficult words stands out as one good way to make an assignment. Another method almost as effective is the development of a background for understanding the material to be studied. This background of understanding is of undoubted assistance in comprehending the meaning of the material studied.

Since these assignments were tried out upon Grades V, VII, and IX, opportunity was offered to study the differential effect of the assignment upon each grade.

TABLE VII COMPARATIVE EFFECTIVENESS OF THE FIVE METHODS OF MAKING THE ASSIGNMENT ON FIFTH, SEVENTH, AND NINTH GRADE PUPILS

	Grade V	Grade VII	Grade IX
Word knowledge	C	C	C
Sentence meaning	E	E	B
Paragraph meaning	D	E	B

Teaching the meaning of the difficult word (Situation C) caused the greatest gain in word knowledge for each of the three grades tested. Making pupils aware of the general meaning of the materials (Situation E) caused the greatest gain in sentence meaning on the part of fifth and seventh grade pupils while with the ninth grade it was the assignment of pertinent questions (Situation B). In developing paragraph meaning, making the pupils aware of the personal value of materials appears to be best in the fifth grade, while making pupils aware of the general background of the material studied is most successful in the seventh grade.

While the gains in the experimental situations over the control situations are not statistically significant, except in two cases of word knowledge (Situations A and C), yet our data show strong indications of the influence of the assignment. The consistency of the gain, the fact that each type caused a noticeable gain on the part of one or two of the grades examined, the brief time for acquainting the pupils with the specific procedure used in the assignment, and the large per cent of a year's gain in many cases, are evidences of the importance of the assignment.

REFERENCES

Book, W F.: The Development of Higher-Orders of Perceptional Habits in Reading *Jour Educational Research*, Vol XXI, No 3, March, 1930.

Crawford, C. C. and McDonald, Lois P. "Modern Methods of Teaching Geography." Boston: Houghton Mifflin Company, 1929, pp 238-240

Jordan, A. M.: *Educational Psychology* Henry Holt & Co, 1928, Chap IV.

Monroe, W S, "*Directing Learning in the High School*" New York: Doubleday Doran & Co, 1928, pp. 418 420.

Monroe, W. S. and Mohlman, Dora K.: Training in the Technique of Study. *University of Illinois Bulletin*, Vol. XXII, No. 2, *Bureau of Educational Research Bulletin* No. 20, Urbana, 1924.

Odell, Charles: The Assignment of Lessons. *University of Illinois Bulletin*, Vol. XXIII, No. 7, Oct. 19, 1925, p. 4.

Sears, J. B.: "Classroom Organization and Control." Boston. Houghton Mifflin Co., 1920, p. 220.

Stone, C. R. "Supervision of the Elementary School." Boston. Houghton Mifflin Co., 1929, p. 311.

Waples, D. "Procedure in High School Teaching." Macmillan Co., 1921.

FACTUAL MEMORY OF SECONDARY SCHOOL PUPILS FOR A SHORT ARTICLE WHICH THEY READ A SINGLE TIME

ALFRED G. DIETZE AND GEORGE E. JONES

University of Pittsburgh

(Concluded from November Issue)

FORGETTING

Comparison of Group Averages. Table VII repeats the averages presented in Tables I to VI[1] in such a manner as to reveal the influence of lapse of time to bring about forgetting of an article read a single time. It is clear from these figures that with passage of time pupils remember fewer and fewer of the facts contained in the articles. Grade by grade and article by article eighty-four comparisons may be made between successive intervals. With six exceptions, only one of which is large enough to be significant, the averages of any given interval are smaller than those of the intervals preceding. This is noteworthy in view of the fact that the figures are based on groups of varying abilities.

Table VIII gives the A-score averages for the combined groups.[2] Average scores decrease consistently with time in the figures of this table. The conclusion seems warranted that, within the limits of the present study, the greater the interval between the reading of an article and the subsequent attempt to recall its content the greater is the forgetting for the facts contained.

Results of Parallel Groups. The above findings were checked for accuracy by using the method of equivalent groups. One hundred cases were selected from each of the six experimental groups on the basis of A-scores in immediate memory in such a manner that the six resulting groups were exactly equivalent and that the A-scores of each were normally distributed. Since 56.7 happened to be the A-score average of the combined grades, we selected 56.0 as the mean and 8.6 as the standard deviation of the new parallel groups and then computed the theoretical frequencies for six class intervals of four points each above and below the mean in a normal curve with mean and standard deviation as above and area equal to one hundred The

---

[1] See first half of this article in previous issue.

[2] It should be recalled that the A-score average represents the work of all groups taken together in terms of the material Arkwright

TABLE VII AVERAGE MEMORY AFTER DIFFERENT INTERVALS[1]

Recall interval	Grades						
	VII	VIII	IX	X	XI	XII	All
Germans							
Immediate	54.8	57.0	67.3	66.6	74.0	75.8	64.8
One day	49.0	55.1	51.2	58.8	59.3	71.8	56.7
Fourteen days.	37.0	34.3	38.8	44.2	46.7	51.3	42.0
Thirty days	32.0	*41.3*	32.1	32.1	40.1	39.3	30.4
One hundred days		20.4	26.5	31.4	35.9	33.3	32.4
Radium							
Immediate	67.3	60.0	72.1	70.0	77.0	84.3	73.4
One day	47.7	55.0	50.1	50.3	61.0	72.0	58.7
Fourteen days	40.7	45.2	44.0	40.8	50.6	40.1	43.7
Thirty days	37.2	47.7	37.1	*42.2*	42.1	37.0	38.3
One hundred days	27.3	31.0	*38.6*	36.0	35.1	*39.1*	33.8
Arkwright							
Immediate	43.0	50.0	60.3	57.4	62.7	61.8	51.6
One day	33.5	43.4	42.4	49.7	55.2	63.3	46.4
Fourteen days	*37.5*	28.8	32.0	41.4	32.3	32.0	33.1
Thirty days	25.0	27.6	25.1	*37.8*	31.5	31.3	30.3
One hundred days.	20.5	22.0	21.0	28.7	23.0	22.6	23.4

[1] Italicized figures show scores that are higher than preceding interval

TABLE VIII.—AVERAGES MEMORY AFTER DIFFERENT INTERVALS—A-SCORE

Recall interval	Grades						
	VII	VIII	IX	X	XI	XII	All
Immediate	50.0	51.3	55.5	50.0	62.3	65.2	56.7
One day	40.0	45.3	40.0	47.9	53.2	61.1	48.0
Fourteen days	30.5	34.8	35.0	36.1	39.0	36.8	36.3
Thirty days.	26.5	33.2	29.5	35.7	34.0	33.8	32.4
One hundred days.	25.5	27.0	28.5	30.5	29.6	31.2	30.0

hundred cases of each group were then taken to fit these frequencies. An attempt was made to keep the number from each grade approximately the same.

The reader should note that this method makes possible the comparison of different intervals on the basis of a given material even though a different group was used for each interval; *e.g.*, in the case of Radium we may compare the results of Group I or II in immediate memory with *Group V in memory after one day*, *Group VI in memory* after fourteen days, Group IV in memory after thirty days, and Group III in memory after one hundred days The six groups are almost exactly alike in terms of A-scores in immediate memory.

The results for each material are presented in Table IX. The reader should note *especially the close correspondence* between the A-score column of the present table with the last column of Table VIII. Such close correspondence between two methods of analysis is an indication of the validity of the methods.

TABLE IX AVERAGE MEMORY OF PARALLEL GROUPS AFTER DIFFERENT INTERVALS[1]

Interval	Radium	Germans	Arkwright	A-score
Immediate	68 0	67 7	50 0	56 0
One day	61 3	40 4	47 8	47 8
Fourteen days	45 0	37 7	34 0	36 0
Thirty days	41 2	36 0	26 2	32 4
One hundred days	35 4	32 7	22 4	28 7

[1] Each average is based on one hundred cases, the A-scores on three hundred.

Rate of Forgetting A curve of forgetting may be plotted either from the data of the last column in Table VIII or from those of the last column in Table IX. Such a curve would be representative of the forgetting of pupils in Grades VII to XII for the article "Arkwright" after a single reading The data of Table VIII are plotted in the forgetting curve of Fig 3 The curve is begun at sixty-three per cent, since this is approximately the amount of the test which pupils in these grades can answer when they are permitted to look up the answers in the text,—*i.e.*, sixty-three per cent of the article "gets across" in a single reading (see *supra*. paragraph entitled, "Comprehension Difficulty and Carelessness"). To enable the reader also to determine how much of what actually "gets across" in a single reading is remembered at each of the intervals, a second scale is provided to the right of the graph which begins the curve at one hundred per cent. The asymptote drawn in represents the estimated point of complete forgetting for the material, which is taken at six per cent since it was found that pupils could answer this amount of the test without previously having read the article (see *supra*, paragraph entitled "Previous Knowledge").

Summarizing and interpreting the data presented, we find that pupils in the grades studied are able to answer an average of sixty-three per cent of the questions on Arkwright when they are permitted to look up the answers in the article, that immediately after the reading they remember an average of 56.7 per cent; after one day an average of 48.0 per cent, after fourteen days an average of 36.3 per cent, after thirty days an average of 32.4 per cent; and after one hundred days an average of 30.0 per cent. The immediate memory score represents an average loss of 43.3 per cent of the facts presented in the article read; of this, the difference between one hundred per cent and sixty-three, or thirty-seven per cent, is due to failure to comprehend and to careless working habits, and the remainder, 6.3 per cent, to forgetting during the interval of reading the article and taking the test. If we compute the per cent remembered of that which actually "gets across," we obtain scores as follows: Immediate memory 90.0 per cent, memory after one day 76.2 per cent, memory after fourteen days 59.3 per cent,

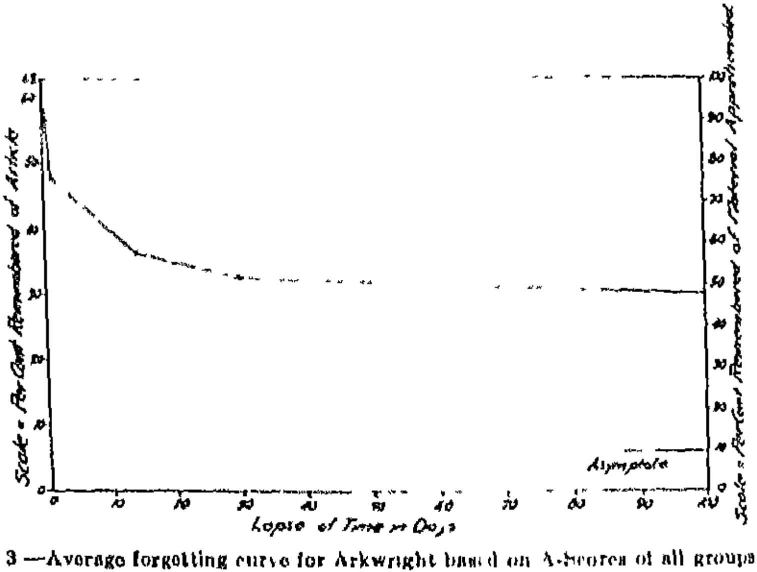


FIG. 3.—Average forgetting curve for Arkwright based on A-Scores of all groups participating in the experiments

memory after thirty days 51.7 per cent, and memory after one hundred days 47.6 per cent.

We find here the same principle discovered by Ebbinghaus,—*viz.*, *that in forgetting there is a rapid initial loss and a more gradual subsequent loss. We find also the principle corroborated that meaningful material is forgotten less quickly than nonsense syllables; and we may add that facts learned in a single reading of an article of the difficulty of Arkwright are, in the longer intervals at least, forgotten less rapidly than in meaningful material in the form of poetry learned to the degree of one correct and unhesitating recital.* In Table X we present data taken from three classic studies of forgetting[1] and compare them with our results. A study of the table shows that the amounts forgotten of what is actually apprehended of Arkwright are less, in general, than for the other materials, especially after intervals of more than one day. From the position of the asymptote in Fig. 3, it is also apparent that the point of complete forgetting is still far off even after one hundred days, and, from the shape of the curve, that it would take a long time to reach this point, if it were ever reached.

TABLE X MEMORY AFTER DIFFERENT INTERVALS

Comparison of Results of Ebbinghaus, Radossawljewitsch and Finkenbinder with Present Study

Interval	Nonsense syllables			Poetry	Factual
	Ebbinghaus	Radossawljewitsch	Finkenbinder	Radossawljewitsch	Dietze
Immediate (20 minutes)	58	80	23[2]	90	90
One day	34	68	58	70	76
Fourteen days	22[1]	41	47[2]	30	59
Thirty days	21	20	41[2]	24	52
One hundred days	?	4[1]	38[2]	?	48

[1] By *interpolation in author's tables*

[2] By entry in Finkenbinders tables of 10 Log

---

[1] Ebbinghaus, H *Ueber das Gedaechtnis.* Leipzig, 1885 translated by H A Ruger and E Bussenius, *Memory* Bureau of Publications, Teachers College, Columbia University, New York, 1913.

Radossawljewitsch, P R, *Das Fortschreiten des Vergessens mit der Zeit* Druck der Dieterich'schen Universitaet, Buchdruckerei, Goettingen, 1905.

Finkenbinder, E O The Curve of Forgetting. *American Journal of Psychology,* Vol XXIV, 1913, pp 8–32.

The conclusion is reached on the basis of the above analysis, that (*a*) forgetting of factual material read a single time follows the same general law of rapid initial loss and more gradual subsequent loss as was found by other investigators for rote memory, (*b*) forgetting of factual material proceeds more slowly than forgetting of nonsense syllables, (*c*) forgetting of factual material proceeds more slowly than forgetting of meaningful material learned by rote, and (*d*) it takes much longer to approximate a point of complete forgetting for factual material than it does either for nonsense syllables or meaningful material learned by rote. It should be remembered that the above comparisons are between memory for facts apprehended in a single reading of an article as measured by a recognition test and verbatim memory for materials learned by rote to the point of one perfect and unhesitating reproduction.

CORRELATIONS BETWEEN MEMORY INTERVALS

Introduction.—The problem of forgetting will be still further clarified by an analysis of the relationship existing between the scores of the same individuals after different intervals, *i.e.*, of the question, "Do pupils rank in the same order, in memory after different time intervals?" This can be done most expeditiously by the correlation method.

Several workers have investigated the correlation between immediate and delayed memory. Among these, Thorndike found a correlation of .9 between immediate recall for twelve unconnected words and delayed recall after twenty-four hours.[1] Gates found correlations of .73 to .89 between immediate and delayed recall of nonsense syllables and of biographies. His subjects were children in the elementary school.[2] Bassett reports average correlations of .838 between immediate recall of history courses and recall after four months, .797 between immediate recall and recall after eight months, .830 between immediate recall and recall after twelve months, .810 between immediate recall and recall after sixteen months, .935 between recall after four and eight months, .915 between recall after eight and twelve months, and .961 between recall after twelve and sixteen months.

---

[1] Thorndike, E. L.: The Relation between Memory for Words and Memory for Numbers, and the Relation between Memory over Short and Memory over Long Intervals. *American Journal of Psychology*, Vol. XXI, 1910, pp. 487–488.

[2] Gates, A. I.: Correlations of Immediate and Delayed Recall. *Journal of Experimental Psychology*, Vol. IX, 1918, pp. 489–496.

She concludes that time and forgetting do not greatly reduce the relative standing of pupils in retention of history learned in school.[1]

The *correlations obtained by the authors cited are, in general,* fairly high. Indeed, Thorndike makes the statement, "The relation between retention of the effects of an experience for one or two minutes and their retention for one or two days thus seems to be one of the closest yet measured in human nature." However, we must remem-

TABLE XI.—CORRELATIONS BETWEEN FACTUAL MEMORY AFTER VARIOUS INTERVALS

r, Variables	r	PE	r corrected for attenuation	*N*
Immediate Radium; one day Germans	75	017	85	293
Immediate Germans; one day Arkwright	78	016	83	284
Immediate Arkwright; one day Radium	73	010	80	280
Average: Immediate; one day	75		83	
Immediate Radium; fourteen days Germans	61	026	66	273
Immediate Germans; fourteen days Arkwright	60	030	51	289
Immediate Arkwright; fourteen days Radium	53	020	58	282
Average. *Immediate; fourteen days*	55		60	
Immediate Radium; thirty days Arkwright	48	032	53	208
Immediate Germans; thirty days Radium	50	030	54	288
Immediate Arkwright; 30 days Germans	58	030	62	225
Average: Immediate; thirty days	.52		50	
Immediate Radium; one hundred days Arkwright	36	039	39	223
Immediate Germans; one hundred days Radium	48	032	52	201
Immediate Arkwright; one hundred days Germans	50	030	54	204
Average: Immediate; one hundred days	45		48	
Fourteen days Radium; thirty days *Germans*	58	032		192
Fourteen days Germans; thirty days Arkwright	58	030		223
Fourteen days Arkwright; thirty days Radium	50	033		241
Average: fourteen days; thirty days	55			
One day Radium; one hundred days Germans	33	056		116
One day Germans; one hundred days Arkwright	33	062		93
One day Arkwright; one hundred days Radium	38	053		119
Average: *one day; one hundred days*	35			

---

[1] *Bassett, S. J.* "Retention of History in the Sixth, Seventh and Eighth Grades with Special Reference to the Factors That Influence Retention." Johns Hopkins Studies in Education No. 12. The Johns Hopkins Press, Baltimore, 1928, pp. 22–24.

COMMENTS

1. Since the type of material which pupils read and study markedly influences the amount which they apprehend and remember, methods should be devised for scientifically matching the difficulty of text-books to the ability level of the pupils for whom they are intended

2. More research is needed to determine the factors inherent in material read and in the mental make-up of pupils which determine the efficiency of memory.

3. It is obvious that a single reading of lesson assignments is an inefficient method of study Studies are needed to discover more effective study techniques. The influence of repeated readings, summarizing, reading to find answers to questions and to solve problems, are suggested problems A veritable mine of similar problems are suggested throughout the pages of G A Yoakam's recent book, *Reading and Study*[1]

4. Other interesting fields of experimentation are suggested by statements nine, ten, eleven and fourteen of the foregoing summary

---

[1] The Macmillan Co , New York, 1928

EFFECT OF ORDER OF PRESENTATION ON THE RECALL OF PICTURES

DANIEL D. DROBA

Ohio State University

Results described in this report were obtained in connection with an experiment on the effect of printed information on memory for pictures.[1] One hundred twenty students attending the University of Chicago were used for subjects. For memory material twenty uncolored postcard pictures of paintings in the Chicago Art Institute were used throughout as follows:

1. A Woman in Gray, by William Orpen, English, 1878- (14)
2. The Fountain, by John S. Sargent, American, 1856-1925 (10)
3. Joan of Arc at the Court of Chinon, by Maurice Boutet De Monvel, French, 1850-1913 (13)
4. Arab Scouts, by Adolph Schreyer, German, 1828-1899 (20)
5. The Music Lesson, by Gerard Terburg, Dutch, 1617-1681 (10)
6. The Young Duchess, by John Faed, Scotch, 1820-1902 (10)
7. Marshes in the North of Holland, by Eugene Jettel, Austrian, 1835-1901 (5)
8. The Potato Harvest, by Ludwig Knaus, German, 1829-1910 (11.5)
9. Democritus, The Laughing Philosopher, by Jusepe Ribera, 1588-1656 (17)
10. Portrait of Helena Dubois, by Anthony Van Dyck, Flemish, 1599-1641 (18)
11. Rembrandt's Father, by Rembrandt Harmensz Van Ryn, Dutch, 1606-1669 (2.5)
12. The Wrestler's Challenge, by Mihaly Munkacsy, Hungarian, 1846-1900. (11.5)
13. Night in the Garden of Gethsemane, by Lucas Cranach, German, 1472-1553 (7)
14. A Wrangle Over Cards, by Jean David Col, Belgian, 1822-1900 (1)
15. The Assumption of the Virgin, by Domenikos Theotokopoulos, Spanish, 1547-1614 (4)
16. The Repentance of Simon Peter, by Jean Charles Cazin, French, 1841-1901 (6)
17. Pilots, Puerto De Los Pasajes, Spain, by Frank W. Brangwyn, English, 1867- (15)
18. He That Is Without Sin Among You, by Benjamin West, American, 1738-1820 (9)
19. The Water Mill, by Meindert Hobbema, Dutch, 1638-1709 (8)
20. In Holland Waters, by Paul Jean Charles Clays, Belgian, 1819-1900 (2.5)

---

[1] D. D. Droba, Effect of Printed Information on Memory for Pictures. *Museum News*, Sept., 1929; *Psych. Abstracts*, No. 94, 1930.

Each picture was fastened on a white card and the information pertaining to the picture was typed either at the bottom or to the right of the picture. Six types of information were used ranging from no information to an information including the title of the picture, the last name of the painter, and a sentence about the painter.

The pictures were shown individually to every subject, the exposure time being fifteen seconds for every picture. Immediately after all the pictures were presented the subject was asked to describe on a separate card all the pictures he was able to recall. Twenty subjects were used for each of the six conditions of information.

For studying the relation between the order of presentation and frequency of recall two methods were used. The first was the rank correlation method. Hence, we ascertained the number of times a picture was recalled by the one hundred twenty subjects. The pictures were then ranked on the basis of the frequency of recall. The most frequently recalled picture was put first, the next most frequently remembered picture was put the second, and so on (rank numbers of recall are given in the parenthesis after each picture in the list). The rank correlation between the frequencies of recall and the order of presentation was found to be −.61. This negative correlation tends to show that the earlier a picture is presented the less likely it will be recalled.

The second method used was the graphic method. The twenty pictures were divided into two groups: The ten pictures presented first and the ten pictures presented second. The averages of the number of pictures recalled under each of the six conditions of information was calculated. These averages, together with the total averages for all the one hundred twenty subjects, were tabulated in Table I.

TABLE I

Pictures	I	II	III	IV	V	VI	Total
a. First ten.	12.0	11.4	12.5	11.4	10.0	11.8	11.8
b. Second ten.	15.3	17.1	14.9	14.0	12.0	15.1	14.8
Subjects	20	20	20	20	20	20	120

Under each condition of information, designated by Roman numbers from I to VI, a higher average was found for the second half of the series of pictures. The difference between the total averages for all the one hundred twenty subjects is .40 which is a confirmation of the above finding that pictures presented in the second half of the series were recalled more frequently, on the average, then pictures presented in the first half of the series

Figure 1 graphically represents these averages plotted against the amount of information Curve *a* represents the first ten pictures and

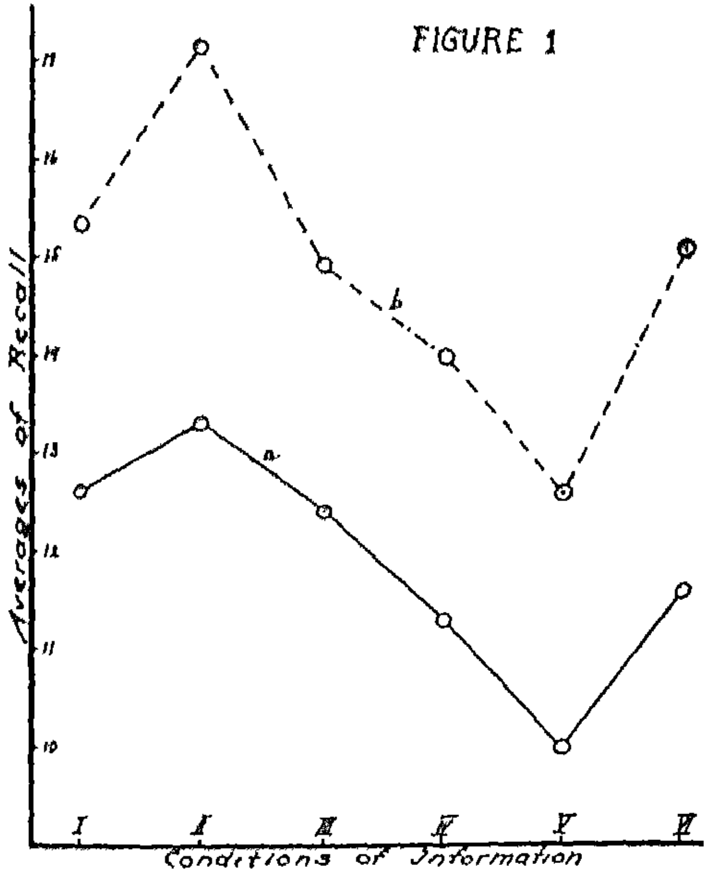
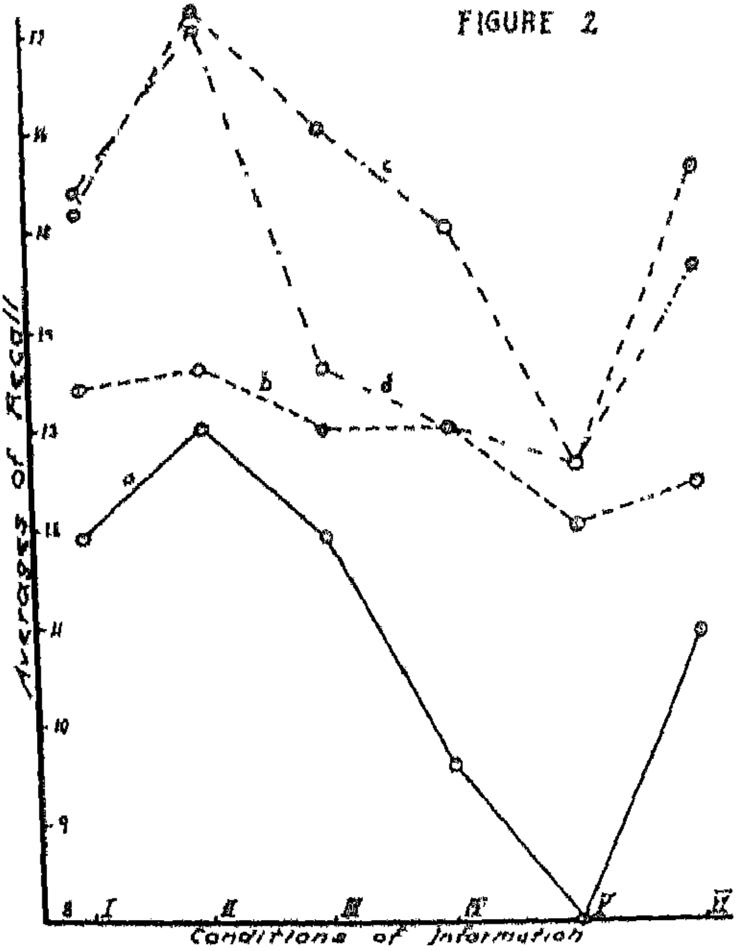


FIGURE 1

curve *b* the second ten pictures. The difference between the two curves is readily seen. *b* is much higher in all conditions than curve *a*. This means that pictures located in the second half of the series were consistently recalled more times than those in the first half of the series

The two halves of the pictures were then again divided into two equal subgroups so that we obtained four groups of five pictures each. Similarly, the averages of recall were calculated for each group under each condition, and tabulated in Table II.



The total averages for all the one hundred twenty subjects are also included in the table. According to the total averages (the last column) pictures in the third quarter were recalled the most frequently. Pictures located in the last quarter come next, the second group the next, and the first group is the last of all.

Data shown in Table II were plotted in Fig. 2. The highest curve is curve *c* representing the third group of five pictures which indicates that the best position is the third quarter in all except the first *and the fifth conditions*. In the latter condition it is equal with the last quarter. Curve *d* representing the last quarter drops down especially in the medium conditions. Curve *b* representing the *second quarter is different from the first two especially under* the end conditions.

TABLE II

Pictures	I	II	III	IV	V	VI	Total
a. First five	11.9	14.0	11.9	9.6	8.0	10.0	10.8
b. Second five	13.8	14.8	14.0	13.0	12.0	12.4	12.0
c. Third five	15.2	17.2	16.0	15.0	12.6	15.6	15.2
d. Fourth five	15.4	17.0	14.6	13.0	12.6	14.0	14.3
Subjects	20	20	20	20	20	20	120

Pictures located in the first quarter and symbolized by curve *a* were decidedly less frequently recalled than pictures placed in the other three quarters. This fact is clearly evident from the figure.

An additional fact may be noted from both figures which is not strictly related to our problem. The relation of the amount of information and frequency of immediate recall is consistent for all the groups of pictures: the halves and the quarters. All curves first *rise to the second condition, then they drop down to the* fifth condition, and again rise for the sixth condition. The meaning of this relation was substantiated in a separate article.[1] However, the *consistency of the relation is being mentioned here for the* first time.

The following conclusions may be drawn from this experiment.

1. The *rank correlation between the order* of presentation and the frequency of immediate recall seems to indicate that priority of presentation is not conducive to such recall.

2. In all the *six conditions of information* the position of a picture in the second half of the series is more favorable to immediate recall than in the first half.

---

[1] *Ibid.*

3. Pictures located in the third quarter of the series were recalled the most frequently of all.

4. Position in the first quarter is decidedly less favorable to immediate recall than positions in the other three quarters.

5. Effect of information on memory for pictures, described in a previous article, was found to be consistent for all the above mentioned positions.

THE INFLUENCE OF AN AUDIENCE UPON RECALL

CLARA BURRI

Graduate Student, University of Chicago

THE PROBLEM

Introspective reports from many observers indicate that the presence of an audience frequently has a detrimental effect upon performance. Many individuals relate that the presence of others causes them many difficulties of behavior especially in speech and mnemonic responses.

Sometimes we find that a student comes to school, seemingly unprepared, yet he maintains that "He did learn his lesson, but *that he just could not remember when others were watching.*" Every teacher is familiar with this or similar statements and he frequently considers it as a mere excuse for not having studied. This investigation is an attempt to study experimentally some of the fundamental phases of the above problem. Our task is to discover whether an individual who has learned certain memory material, and is then required to recall that material before a group of people, will as a consequence of the presence of that audience, be able to recall as readily as when he is alone.

Few people have considered the influence of a group on an individual's efficiency to recall when learning did not take place in that group. Elkine[1] is the only author who reports any data concerning group influence on recall or reproduction. He reports that memory is better in the group. Forty children were given one presentation of a series of one-syllable words and were then called upon for a written reproduction immediately after the presentation, and another one day later. First, the individuals went through the above stated procedure alone, and later as members of a group. Elkine explains his results in terms of "rivalry and the influence of suggestibility due to contact and mere presence of other people." Elkine's study differs from the one reported in this paper in one important aspect. He studied the influence of a group on learning and memory when the subject was a member of the group. The results obtained may therefore be explained as being due to rivalry. In this study particular care was taken to eliminate, so far as possible, the factor of competition, and hence the subjects were tested when recalling before people who were not engaged in any comparable activity.

---

[1] Elkine, F. Influence du groupe sur la mémoire. *J. de Psychol.*, 1927, pp. 827–830.

METHOD AND PROCEDURE

Subjects. Sixty students, nineteen men and forty-one women, from the classes in introductory psychology, and from other departments of the University of Chicago, participated in the experiment.

Conditions. The subjects were divided at random into three groups of twenty each. Group one acted as control, learning and recalling under the normal condition, with only the experimenter present. Group two learned under the same condition as did group one, but they recalled and relearned before an audience consisting of four persons who sat in the experimental room without paying any attention to them. The individuals of group three also learned alone, and then recalled and relearned before an audience. This audience however attentively watched and kept track of their responses.

Audiences. The audiences consisted of two men and two women, all graduate students or majors in psychology. They were selected with a view to their prestige in the eyes of the subjects, and as far as possible consisted of people who were not acquainted with the subjects. No attempt was made to keep the same individuals for spectators throughout the experiment. As has already been suggested two kinds of audiences were employed: One in which the members paid no attention to the subject, but were merely present in the experimental room: Another, an attentive audience, the members of which listened to the subject and checked his responses on a slip of paper which contained the lists of words in their correct order.

Material. The material to be learned consisted of a list of fifteen pairs of words which had been selected at random from a standard dictionary. In constructing the list care was taken that no letter should occur twice in a given pair. The words were exposed by means of a hand-operated memory drum, a standard instrument in the laboratory of the University of Chicago. Other materials used were cards three to five inches with the response words copied on non-corresponding order with the stimulus words of the test series, for recording the responses. A stopwatch served for timing the exposure of the paired associates.

Procedure. The paired associate list was presented once, allowing three seconds for each exposure, after which the subject was tested for recall. In the test or recall session the word was presented alone and in a different order from that in which it had originally appeared, and the subject had to give the proper response word within the given time of three seconds. The recall presentations were continued until the subject was able to respond correctly to every stimulus word on the list. The criterion of learning was one complete recall. A subject learned the list of paired associates and then recalled and relearned them twenty-four hours later either before the experimenter alone, or in presence of an audience, attentive or non-attentive, depending upon the particular condition desired.

Instructions. At the beginning of the experiment each subject was required to read a carefully prepared and typewritten instruction sheet, after which he was questioned briefly to make certain that he understood the general nature of his task. At the same time he was also told under what condition he was expected to recall, and relearn on the following day. Thus every subject knew before he started to learn, whether an audience, and if so, what kind of an audience, was to be present for recall.

The only subjects given other instructions on the recall day were those who were to face an attentive audience. To them, the request was made to "respond distinctly and loudly, so that every person in the room can understand." This extra order was given to enhance the effectiveness of the attentive audience and thus to further increase the difference between the non-attentive and attentive audience.

Treatment of Data. The amount of retention is stated in terms of (1) number of words recalled, (2) number of trials required to relearn; and (3) the saving score, which was obtained by the following formula:

$$1 - \left(\frac{\text{Trials to relearn}}{\text{Trials to learn}}\right)100$$

The difference in efficiency of recall between the several conditions is determined statistically and the results are stated in terms of the numerical differences, the averages, the probable errors of the means, the difference between the means, and the probable error of their differences.

RESULTS

Table I gives the total and mean number of words recalled, trials required to relearn and the saving score for the different conditions. If we analyze the data it appears that the results in terms of recall are most clear cut, while those in terms of relearning and saving, although less large, still show a definite tendency. The efficiency of retention as measured in terms of recall, relearning, and saving, is greatest for no audience. In the case of an audience, and an attentive audience, the numerical scores for retention are considerably smaller than those for no audience, while on the other hand they are about equal for the two kinds of audiences. The statistical significance of the findings is shown in the later tables. These results also throw light upon a fundamental principle of reproduction. They indicate that any stimulus-response situation must be considered as a functional whole consisting of a constellation of numerous organismic and environmental factors. Any change in the situation will affect the synthesis of the whole. If a behavior pattern is once formed it may be reinstated best in a similar situation to that in which it was originally acquired. In our problem not only the total organization present at the time of learning, but the constellation of organismic and environmental factors at the time of recall were equally important in facilitating or inhibiting reproduction.

TABLE I. THE EFFICIENCY OF RETENTION IN THE DIFFERENT CONDITIONS

Condition	Recall, number of words recalled	Relearning, trials to relearn	Saving: percentage of learning time saved
No audience			
Total	269	18	1561
Average	13.45	.90	78.05
Audience			
Total	237	33	1387
Average	11.85	1.65	69.35
Attentive audience			
Total	235	32	1417
Average	11.75	1.60	70.85

Table II shows the mean number of words recalled, of trials required to relearn, and of the saving per list for each of the two conditions, no audience and audience; the probable errors of the two means from the three criteria and the probable errors of their differences. There is a difference between the two conditions in all of the three criteria. The one in terms of recall is almost four times its probable error. The differences in the case of the relearning and saving scores is not so great, yet they indicate a consistent tendency. This is especially remarkable since these scores are not direct measures of recall.

TABLE II. THE DIFFERENCE BETWEEN NO AUDIENCE AND INATTENTIVE AUDIENCE

Condition	Recall, syllables recalled	Relearning, trials to relearn	Saving, percentage of learning time saved
No audience			
Mean	13.45	.90	78.05
PE	.28	.12	4.14
Audience			
Mean	11.85	1.65	69.35
PE	.34	.21	4.07
Difference between the two means	1.60	.75	8.70
PE	.44	.24	6.21

In Table III we find the mean number of the words recalled, of trials required to relearn, and of the saving per list for each of the two conditions, no audience and attentive audience, the probable errors of the two means from the three criteria, and the probable error of their differences.

TABLE III. THE DIFFERENCE BETWEEN NO AUDIENCE AND ATTENTIVE AUDIENCE

Condition	Recall, syllables recalled	Relearning; trials to relearn	Saving, percentage of learning time saved
No audience			
Mean	13.45	.90	78.05
PE	.28	.12	4.11
Attentive audience			
Mean	11.70	1.60	70.85
PE	.17	.16	4.23
Difference between the two means	1.75	.70	7.20
PE	.32	.20	5.91

As before we find that the subjects with audience give poorer results. The difference between no audience and attentive audience is even larger than is the difference between no audience and inattentive audience. Again we get the largest divergence for recall with a disparity between the two averages of 1.75 and a probable error of the difference of .42, which is less than one-fourth of the actual difference. This indicates a high degree of significance.

The results become more striking when we consider that the differences exist in spite of the fact that the subjects took, on an average, longer to learn the paired associates when they were anticipating an audience, than when expecting to recall only before the experimenter. We generally find that a longer learning period is followed by better recall. One can scarcely avoid the inference that the lengthened learning period is an outcome of the expectancy of an audience. The discussion of this topic, however, will have to be left until later.

TABLE IV.—THE DIFFERENCE BETWEEN INATTENTIVE AUDIENCE AND ATTENTIVE AUDIENCE

Condition	Recall, number of words recalled	Relearning, trials to relearn	Saving; percentage of learning time saved
Audience			
Mean	11.85	1.65	69.35
PE	.14	.21	4.67
Attentive audience			
Mean	11.70	1.60	70.85
PE	.17	.16	4.23
Difference between the two means	.15	.05	1.50
PE	.22	.26	6.30

Table IV gives the mean number of words recalled, of trials required to relearn and of the saving per list for each of the two conditions, inattentive and attentive audience, the probable errors of the two means from the three criteria and the probable errors of their differences. There is practically no disparity between the results for the two kinds of audiences.

It seems that the kind of audience is not very important as regards its influence on recall. However, our results may be a function of the particular situation, and if that should be the case, other group arrangements might have different consequences. Therefore no general conclusions can be made. From the viewpoint of the organization of the audiences further observations and studies would be desirable.

Table V contains the data on the learning of the paired associates. In all cases learning took place with only the experimenter present. The sole difference between the several learning conditions was that, in the first, the subjects did not expect an audience, while in the other two, they did. Naturally we would expect the learning efficiency to be about equal throughout the experiment, or at least not to vary to a greater degree than could be explained as due to chance. Yet we find that learning was on the average considerably easier when the subjects did not expect an audience. The difference in trials required to learn the series of paired associates when expecting no audience and when anticipating an audience is a total of 28 trials or a mean of 1.40 trials. The probable error of the difference of the two means is .45. The actual disparity is a little over three times its probable error, indicating significance. There are no significant differences for the two different kinds of audiences.

TABLE V. THE EFFECT OF EXPECTANCY OF AN AUDIENCE FOR RECALL, AT THE TIME OF LEARNING, ON THE RATE OF LEARNING

Condition	Total trials to learn	Average trials per list	PE
No audience	96	4.80	.32
Inattentive audience	124	6.20	.32
Attentive audience	123	6.15	.31
Difference between no audience and attentive audience	28	1.40	.45
Difference between inattentive and attentive audience	0	0	

The fact that the people needed more trials to learn the material when they were expecting an audience may be explained as being due to difference in anticipation, and integration with reference to the learning situation. Anticipating an audience seemed to have been inhibitive to the learning process. The subjects' reports clearly indicate that their knowledge of having to recall before spectators caused some uneasiness. They appeared to learn with the intention of knowing the words well, and of being able to tell them to somebody. Such an attitude was not recognizable with the subjects who did not look forward to an audience; these subjects appeared to learn without any definite end in view. These are the facts as they are before us. We conclude from them that differences in expected conditions of recall may cause corresponding differences in rate of learning.

CONCLUSIONS

Restating the facts as they appear from our investigation we find that:

1. If learning takes place privately, the presence of a group at the time of recall is detrimental to the efficiency of reproduction.

2. The kind of audience is not very important as regards its influence on recall. However, it may be that in this connection the results are a function of the particular situation, and that other group arrangements might have different consequences.

3. The anticipation of an audience for recall, at the time of learning, increases the learning time, and, in spite of this, the amount remembered is less when recall occurs before an audience.

These results suggest that the pupil who complains of not being able to remember well when requested to recite before a class may not be trying to escape responsibility. His inability to recall reveals a fundamental factor of memory, and shows that the situation during the time of learning and during the time of recall are equally important in facilitating or inhibiting the response.

THE RELATIVE INFLUENCE OF VISUAL AND AUDITORY FACTORS IN SPELLING ABILITY

GEORGE W. HARTMANN

Department of Psychology, Pennsylvania State College

A. Background of the Problem

The spelling of the average undergraduate is notorious. One suspects that the great vogue of objective examinations may be partly explained by their power to alleviate eyestrain due to illegible papers, disgust caused by childish syntax, and irritation born of repeated misspellings. These common deficiencies are especially prominent in science courses where the difficulties of mastering a technical vocabulary appall even the *superior student*. *How many instructors are there* who after exposing their charges to a year of its content find them *spelling psychology more like* philosophy, physiology, and philology than like its own true self?

The various explanations which have been advanced for special abilities and disabilities as they manifest themselves in academic or vocational fields are suggestive but hardly convincing. In the case of such a restricted function as spelling, *e.g.*, it is easy to demonstrate that it is not an exclusive property of high or low intelligence, although it is equally certain that it is not as independent of general native ability as some extremists would have us believe. An analysis of the commonest types of errors leads to the suspicion that various sensory *and motor idiosyncrasies are the major determiners of performance*, but *which* ones are crucial is as yet unknown. The current view seems to be that every poor speller must be individually diagnosed and specific remedies prescribed—an obvious consequence of an implicit belief in plural causation.

However, the scientific search for adequate generalizations is not to be dismissed so lightly. Just as it is profitable to know that superior intellects tend to have correspondingly better physiques, emotional control, etc., so we ought to know what the significant characteristics of a good speller are as distinct from a bad one.

Among the many explanations advanced for *poor spelling, perhaps* the most recurrent one is that which attributes it to the non-phonetic *nature of the English language. Were certain sounds uniformly* represented by definite letter-forms thus alone would facilitate correct spelling. The advocates of reformed spelling apparently take this seriously, despite the fact that errors occur with relatively equal frequency in highly phonetic tongues such as German or Spanish. A slightly more plausible variant of this argument is that if a person could pronounce words correctly he would have no difficulty with their orthography. Unfortunately, this fails to account for the fact that many Americans can spell French terms properly but pronounce them in a very unorthodox manner; and similarly for the Frenchman using English. Moreover, most misspellings are substitutions or displacements of letters which by no stretch of imagination can be attributed to faulty pronunciation. The definitive elimination of the pronunciation theory would seem to be an inevitable consequence of the finding that deaf-mutes are relatively better in spelling than in other school subjects.

Other viewpoints have emphasized the rôle of visual perception. Here the commonest opinion is that poor spellers are deficient in visual imagery, but the tachistoscopic span of apprehension has also been considered as the decisive factor.

Where such diverse and contradictory opinions exist, a resort to experimental attack is indicated. To the writer, the problem phrased itself as follows:

Does excellence of auditory or visual perceptive capacity play the dominant rôle in spelling performance? Is spelling ability more closely associated with those reactions mediated by the ear or with those involving the eye? The wording of these questions by implication minimizes the influence of motor impulsions as it is assumed that chance slips of the pen should occur equally often with both groups, and it also sets in the background the influence of such central factors as intelligence, type of imagery, association, etc.

B. THE TESTING PROCEDURE

This investigation began by administering to large samples of each college class from Freshman to Senior (six hundred thirty-six cases in all) a test consisting of fifty selected words appearing in both the Thorndike and Horn Word Books. Upon a prepared sheet, the subjects *printed* each word as it was spoken and its meaning illustrated by a sample sentence. That the test was a reasonable measure of spelling ability is indicated by the reliability coefficient of .81 ± .007, obtained by the split-half technique. For the purposes of this experiment, representative cases chosen from the middle and both tails of the distribution were selected for intensive study. Sixty-three individuals were thus obtained twenty-four of the best spellers, eighteen of the poorest, and twenty-one average ones centering around the median score.

The following test series was administered individually to each of the subjects

Test 1. *Perceptual Span for Meaningful Material* (*Reliability* .74) Upon the middle of twenty small white cards were printed the following unusual words, derived largely from Whipple's well-known information test:

ageratum
Caedmon
cleistogamous
gneiss
guimpe
hemiptera
impetigo
intaglio
metacarpal
peneplain
simony
synecdoche
trephine
trilobite
Weismannism
onomatopoeia
sphygmomanometer
exophthalmic
Abderhalden
La Rouchefoucauld

One-second exposures of the cards were made tachistoscopically; during the intervals (which varied with the writing time) the subject recorded the word he had seen The total number of reproductions correct in every detail constituted the score

Test 2. *Visual Recognition* (*Reliability* .64).--Twenty common words printed on cards were exposed to the subject at the rate of about one per second with instructions to fixate each item Immediately after the original series had been seen, the cards were shuffled with twenty other confusion stimuli. The mixed set was then arranged by the subject into two piles one containing the cards the subject had seen and the other containing those he had not seen Score = 2 × number of errors in both piles subtracted from 40 (For the exact items see the list in the writer's appendix to *Archives of Psychology*, No 100.)

Test 3. *Silent Reading*. Whipple's High School and College Form B was administered in accordance with the time limits and other conditions specified in his directions sheet. Score equals the number of correct answers.

Test 4. Hidden Word Identification (Reliability .88). This consists in underlining a meaningful English word in a page of pied material, the material being embedded in a jumbled nonsense collection. A two-minute time limit was imposed. Score equals the number of words properly underlined.

Test 5. Letter-Digit Substitution Code (Reliability [illegible]). A simple "home made" code using the ten Arabic figures was devised. Time of transcription was measured in seconds.

Test 6. Auditory Memory Span for Digits (Reliability .71). The fourteen number sets from the corresponding section of Whipple's manual were read to the subject in time with a metronome, and recorded by him as soon as heard. Averaging the highest correct reproduction in both sets gave the desired measure of span.

Test 7. Pronunciation (Reliability [illegible]). The thirty two words forming the pronunciation section of the Cross English Test were used. The printed list was shown the subject with instructions to pronounce each word as though he were being graded for purity of speech by a language teacher. The experimenter checked the correct items by listening carefully as the subject proceeded.

Test 8. Auditory Recognition (Reliability [illegible]). A list of ten ordinary words was read at an even rate to the subject with instructions to fixate them. Immediately after this list had been read, a longer list of twenty words containing the ten originals plus ten confusion stimuli, was offered. As each word in the test list was uttered, the subject said "yes" if he had heard it before and "no" if he had not.

The reliability coefficients presented for each of the foregoing measures were obtained by duplicate forms in the case of Tests 2, 4, 5, 8 and by matching alternate items in tests 1, 6, 7; for Test 3 adequate reliability standards were assumed to have been met by the originator.

C. Results

The mean performance on each of the preceding eight tests was separately computed for each group of spellers—the superior, the average, and the inferior. Table I below summarizes the essential data for all classes.

Before proceeding to an interpretation of the table, let it be noted that if spelling ability be conditioned by excellence in visual perception then the "good" group should reveal a definite superiority in Tests

TABLE I.—COMPARATIVE PERFORMANCE OF THREE CLASSES OF SPELLING ABILITY IN VARIOUS VISUAL AND AUDITORY MEASURES

Spelling scores	1 Meaningful perceptual span	2 Visual recognition	3 Silent reading	4 Hidden words	5 Substitution	6 Auditory memory span	7 Pronunciation	8 Auditory recognition	Constants	Class
38.25	9.73	69.16	9.34	13.76	61.01	6.95	17.55	82.65	Mean	Total group
7.32	3.07	13.29	3.06	3.90	8.86	1.88	2.61	7.78	SD	
47.62	12.46	75.63	10.95	14.85	55.52	7.15	19.96	85.94	Mean	Superior
1.26	1.95	11.87	2.82	3.39	6.42	.75	3.75	7.69	SD	
29.15	4.7	2.09	2.05	.88	3.81	.33	1.34	1.4	D/SD$_{diff}$	
38.0	9.76	67.55	9.29	13.85	64.09	7.05	18.62	82.98	Mean	Average
1.95	1.89	13.79	2.62	4.18	8.48	1.22	2.91	6.43	SD	
18.24	4.81	1.15	2.43	1.29	.24	1.05	5.08	2.37	D/SD$_{diff}$	
23.77	6.28	62.36	7.28	12.22	64.75	6.28	14.22	77.92	Mean	Inferior[1]
2.81	2.53	14.27	2.55	3.65	8.43	2.53	2.52	6.81	SD	

[1] The D/SD$_{diff}$ comparison is of course unnecessary with the third group.

1-5 inclusive; but if excellence in auditory reactions be the decisive factor then the good groups should be superior in Tests 6-8 inclusive. The expected positions of both the average and poor groups under either hypothesis may be readily inferred.

The object of Table I is to show which tests meet the conditions of standard reliability in exposing the significant differences between groups representing three degrees of spelling ability. It will be seen at once that very few do so. In the visual series, Test 1 alone discriminates the superior from the average and the average in turn from the inferior. Tests 2, 3, and 4 all fall short of the conventional requirement. Test 5 oddly enough distinguishes the good from the mediocre group, but fails to discriminate between the latter and the inferior.[1]

In the auditory series, no test possesses discriminatory power with the exception of Test 7 which does differentiate the average from the poor spellers.

Apparently Test 1 of the visual group is the only one which can be used to segregate the various classes of spelling talent. This test contains a group of rare words which the subject copies from immediate memory as soon as they have been exposed. According to the evidence which this affords, speed and accuracy of immediate visual perception exert the dominant influence on spelling proficiency.

An attempt was next made to get at the heart of the problem by comparing total standard scores in both the visual and auditory batteries for each of the three ability groups. This comparison was made as follows: The means and sigmas of all subjects in each test were determined, and the mean sigma position of each ability group was separately computed for the visual and auditory series. This is a most laborious operation, but the following table gives the condensation:

TABLE II.—MEAN SD POSITIONS OF THREE SPELLING ABILITY GROUPS IN VISUAL AND AUDITORY TESTS

Visual battery (Tests 1-5 inclusive)			Auditory battery (Tests 6-8 inclusive)		
Superior	Average	Inferior	Superior	Average	Inferior
.55	.00	−.62	.48	.04	.69

---

[1] Since the scores in Test 5 represent times taken, the lowest values designate the best performances.

The reasoning which is intended to apply is: If the superior (or inferior) spellers are such primarily because of high (or low) visual perceptual capacity, then their relative performance should be most marked in the visual series. As a matter of fact, the order of excellence is about the same in both batteries. Good spellers, apparently, can use their ears proportionately as effectively as their eyes, which leaves us uncertain as to which of these senses their prowess should be attributed.

As the method of reliable differences yielded such ambiguous evidence, it was necessary to appeal to further analysis in the hope that this would reveal the true relations which conceivably had been obscured by some kind of compensatory masking. Raw correlations between the spelling scores and the eight tests were therefore computed, and the following coefficients obtained:

TABLE III. RAW CORRELATIONS BETWEEN SPELLING SCORES AND PERCEPTUAL TESTS

Test No.	Description	*r*	PE
1	Meaningful perceptual span	78	.03
2	Visual recognition	.30	07
3	Silent reading	.47	07
4	Hidden words	27	08
5	Substitution	41	07
6	Auditory memory span	15	08
7	Pronunciation	58	06
8	Auditory recognition	43	07

This method seems to check the results presented in Table I for again the test of perceptual span is predominant. Since the *r* between spelling and the data of Test 1 is attenuated by the imperfect reliabilities of the respective scores, Spearman's correction formula was applied, resulting in an *r* of almost .99! It appears that the two tests measure much the same function.

A final attempt was made to build up a regression equation from the three most satisfactory visual tests, *viz.*, Tests 1–3, inclusive. The raison d'être of this was that if it could be shown that admittedly visual factors possess adequate predictive power for degrees of spelling talent, then they must be viewed as causative influences thereof (in Hume's sense). The following calculations are accordingly presented:

$R_{3(124)} = .784$

$x_3 = 1.59x_1 + .62x_2 + .10x_4$ (deviation form)

$X_3 = 1.59X_1 + .62X_2 + .10X_4 + 36.24$ (score form)

$SD_{est\ x_3} = 4.54$

It will be noticed that the multiple correlation is in no way greater than the coefficient of zero order between the first test of our series and the spelling scores. The obvious implication is that this test carries almost exclusively the full burden of prediction.

D. CONCLUSIONS

The findings presented in the previous sections do not readily lend themselves to any clear-cut interpretation, but certain provisional statements appear to be justified. In the first place, spelling ability is no more a function of *general visual perception* than it is of general auditory perception. Nevertheless it does seem to be closely related to the *special reaction* involved in reproducing tachistoscopically-exposed stimuli of a meaningful nature. To use the language of Gestalt, good spellers perceive total configurations of the verbal sort with greater facility than others. Secondly, spelling ability is essentially a central function, with normal subjects peripheral factors of an optical or retinal nature are inconsequential. Thirdly, while the best way to determine spelling ability is still the sample or scale method, it would seem possible to use the tachistoscopic perceptual span technique as an alternative thereto, since this has the advantage of brevity. Fourthly, all the preceding statements should be interpreted in the light of the fact that all the data, especially the correlations, pertain to individuals chosen from each of the major segments of a larger curve.

E. SUMMARY

This study concerned itself with the following question: Is there any form of sensory perception which is the major determiner of spelling proficiency? Sixty-three college students, representing respectively the best, the worst, and the middle levels of performance, were given eight laboratory tests *singly*. These tests fell into two groups: one requiring the use of the visual pathways and the other depending on the receptors for sound.

Statistical treatment of the results showed that the visual battery *as a whole* did not discriminate among the groups any better than the general auditory battery, and *eye tests*; one visual test, however (that of immediate memory span for meaningful visual stimuli), correlated .78 ± .04 with the spelling criterion and served effectively to segregate the three classes of ability.

The major outcome of this investigation is that spelling ability is largely dependent upon *one special form* of visual reaction and not upon general superiority in any sense modality or upon a common integrative capacity.

FURTHER DATA CONCERNING THE EFFECT OF WEIGHTING EXERCISES IN NEW-TYPE EXAMINATIONS

C. W. ODELL

University of Illinois

The authors of several of the early standardized tests devoted considerable labor and care to the determination of supposedly accurate weights, usually on the basis of difficulty, for the separate elements composing their tests. Within a few years, however, a strong tendency to discontinue the use of such weights appeared. Studies by Douglass and Spencer[1] and others showed such high coefficients of correlation between scores based upon such weights and so-called "unweighted" or raw scores, that is, those based upon one point for each element, that they led to the conclusion that the considerable amount of extra work involved in deriving and using weights was unnecessary. In a recent article Corey[2] has presented some data which scarcely support this conclusion. He had each item of a new-type test in educational psychology weighted by six instructors in that subject. From the results he determined the correlations between raw scores and those computed according to these six series of weights, and also the effects of the different weightings upon letter marks into which the test scores were transmuted. Five of the six coefficients with the raw scores ranged from .82 to .88, the other being .90. Approximately one-fourth of the papers were given the same letter mark according to all seven scores, and another fourth each according to six, five, and four of the seven. The marks according to the six instructors varied from those based on raw scores in from twenty-two to forty-nine per cent of the cases.

Because of the considerable disagreement between these results and others previously obtained, the present writer was led to make two studies along the same line. Before proceeding to present his own data, however, he wishes to offer one comment on Corey's investigation It does not appear in the article, but the writer happens to know that five of the six instructors assigned weights of zero to some of the elements, and thus altered the conditions of the experiment. A test which contains some items that do not count in determining the score is for all practical purposes no longer the same test as if all the elements were counted, so that the general effect of this factor was to lower the obtained correlations

---

[1] Douglass, Harl R. and Peter L. Spencer. Is It Necessary to Weight Exercises in Standard Tests? *Journal of Educational Psychology*, Vol. XIV, February, 1923, pp. 109–112.

[2] Corey, Stephen Maxwell: The Effect of Weighting Exercises in a New Type of Examination. *Journal of Educational Psychology*, Vol. XXI, May, 1930, pp. 383–385.

Part of the data which the writer wishes to present are based upon a fifty element multiple answer test with four suggested answers to each element that was taken by sixty-two students in educational measurements The elements in the test were weighted upon five different bases in addition to the raw score These bases will be referred to by subscripts as follows.

1 = raw score one point for each element
2 = percentage of students answering correctly
3 = difficulty as determined through application of normal curve to 2
4 = random distribution of weights one to five
5 = random distribution of weights one to ten
6 = second random distribution of weights one to five

It will readily be seen that the bases of weighting employed in this experiment were such as to give no promise of agreement between the weights, thus making the conditions of the experiment much more unfavorable to securing high correlations between raw and weighted scores than if weights had been assigned according to the opinions of instructors who might be expected to agree to some extent at least. Not only this, but in the case of bases two and three a high negative correlation might be expected, since one was based directly upon how easy the elements were and the other somewhat less directly upon their difficulty The correlations between the various lists of weights are given in Table I

TABLE I COEFFICIENTS OF CORRELATION BETWEEN WEIGHTS OF ELEMENTS IN FIRST STUDY

$r_{23} =$ 62 $r_{34} =$ 03 $r_{45} =$ 21 $r_{56} = .03$
$r_{24} =$ 04 $r_{35} =$ 24 $r_{46} =$ 07
$r_{25} =$ 14 $r_{36} =$ 10
$r_{26} =$ 11

The writer's second study was similar except that instead of the three random assignments of weights, three instructors actually assigned them, and that the weights for ease and difficulty were both taken on the same basis so that the correlation between them was

-1.00. The test in this case was a twenty-two element single-answer test in methods of teaching. The coefficients of correlation between the weights in this case are in Table II.

TABLE II. COEFFICIENTS OF CORRELATION BETWEEN WEIGHTS OF ELEMENTS IN SECOND STUDY

$r_{12} = -1.00$	$r_{23} = .08$	$r_{34} = .77$	$r_{45} = .34$
$r_{13} = .09$	$r_{24} = .07$	$r_{35} = .57$	
$r_{14} = .05$	$r_{25} = .00$		
$r_{15} = .04$			

It will be seen that except in the case of the weightings according to ease and difficulty and of those by the three instructors, the correlations between weights were practically zero. Those given by the three instructors correlated positively, in one case decidedly low, and in the other two only fairly high.

After weights had been determined as described, each student's score was computed according to each of the six plans of weighting. In the first study there were two ways under each, one by totaling the points allowed for all the elements to which he gave the correct answers, and the other by means of the approved formula for multiple-answer tests, score $= R - \frac{W}{N-1}$, which in this case becomes score $= R - \frac{W}{3}$. In this, of course, R stands for the number of right responses, and W for the number of wrong responses. In the second study only one method, the number correct, was employed. The intercorrelations between the scores on each of the six bases are given in Table III. In the first half thereof the first coefficients are for scores based upon correct responses only, and the second for scores based upon the formula.

TABLE III. COEFFICIENTS OF CORRELATION BETWEEN SCORES BASED ON DIFFERENT WEIGHTINGS

First Study

$r_{12} = .98, .97$	$r_{23} = .93, .90$	$r_{34} = .90, .88$	$r_{45} = .96, .96$	$r_{56} = .98, .99$
$r_{13} = .92, .92$	$r_{24} = .97, .96$	$r_{35} = .92, .93$	$r_{46} = .98, .98$	
$r_{14} = .98, .98$	$r_{25} = .96, .96$	$r_{36} = .91, .93$		
$r_{15} = .99, .99$	$r_{26} = .95, .95$			
$r_{16} = .99, .99$				

Second Study

$r_{12} = .97$	$r_{23} = .75$	$r_{34} = .91$	$r_{45} = .99$	$r_{56} = .92$
$r_{13} = .96$	$r_{24} = .91$	$r_{35} = .97$	$r_{46} = .98$	
$r_{14} = .98$	$r_{25} = .93$	$r_{36} = .96$		
$r_{15} = .97$	$r_{26} = .96$			
$r_{16} = .98$				

It will be seen that only in the case of the third basis, that is, the weights determined according to the normal probability curve as applied to difficulty, were the coefficients between the raw scores and the weighted scores lower than 97, and that in half of the other cases in the first study they are 99, and in all the others in both studies 97 or 98 Furthermore, among the various bases of unequal weighting the only cases in which the coefficients dropped markedly below .90 were those in which bases two and three were concerned Such a result would be expected from the fact that there was a fairly high negative correlation between the weights given on these bases. Of the three between the different instructors in the second study, two are decidedly high, 98 and 99, whereas the other one is considerably lower, being only 92

The comment should probably be inserted that the tendency of the correlations and other measures obtained from the results of the first study to show greater agreement than those from the second is undoubtedly due to the fact that the test employed therein was much longer than that used in the second

Following Corey's example, the writer also computed the effect of different weights upon marks This was done in the same way as by Corey, considering the 7 per cent highest scores as *A*'s, the next 24 *per cent as B's, the next 38 per cent as C's, the next 24 per cent as D's, and the lowest 7 per cent as E's.* Table IV gives figures showing the per cents of the pupils receiving the same letter mark according to the various numbers of bases of weighting

TABLE IV PERCENTAGES OF PAPERS RECEIVING SAME MARKS ACCORDING TO DIFFERENT WEIGHTINGS

Number of different bases of weighting	Percentages given same marks	
	First study	Second study
6	42, 30	38
5	30, 32	33
4	12, 18	15
3	10, 11	14

From this table it appears that about 40 per cent of the students would have received the same letter mark according to all six bases of weighting, and about another 35 per cent according to five of the six In other words, in about three-fourths of the cases would either all, or all but one, of the different weightings have given the same mark

The percentages of cases in which the marks assigned according to the different bases except the first differed from it were also determined. These are shown in Table V

TABLE V.—PERCENTAGES OF PAPERS ON WHICH MARKS ACCORDING TO DIFFERENT WEIGHTINGS VARIED FROM THOSE BASED ON RAW SCORES

Basis of weighting	Percentage of marks differing from those according to raw scores	
	First study	Second study
2	10, 23	21
3	26, 29	30
4	13, 21	14
5	11, 19	18
6	6, 18	12

A comparison of the results obtained in the writer's two experiments with those of Corey shows that they indicate much less discrepancy or disagreement between scores assigned according to different bases of weighting. In view of the fact that the tests were so short, one containing only fifty elements and the other only twenty-two, the results are even more significant. To the writer they seem to offer very strong support of the conclusion that for new-type tests there is so little to be gained by unequally weighting the elements that it is not worth the labor of assigning such weights and computing scores from them. The longer the test the more fully does this conclusion hold.

THE HANDWRITING OF INDIANS

THOMAS R. GARTH

University of Denver

In order to ascertain whether or not there are racial differences in native traits we shall have to measure those traits. Let us take the simple matter of differences in handwriting between Whites and Indians. It is believed that there are peculiarities in handwriting which run in families in a race. The question is, do these peculiarities run in races? Is Indian handwriting peculiar to itself and different from that of Whites?

Immediately the problem becomes a large one. So many things need to be considered. There is the question of legibility and speed. Another is the question of esthetic quality. Still another is what has been called by graphologists "characterological" interpretation. Yet another is consideration of the diagnosis of the handwriting of the two races. In the present study we have been compelled to confine our inquiry to the first question, *i.e.*, a Comparison of the Legibility and Speed of the Handwriting of Whites and Indians.

Problem

We have asked ourselves these questions;

1. Do these full- and mixed-blood Indians write as legibly as White children of a like school grade?
2. Do they write with the same speed?
3. Which of the factors under consideration have the greater influence on legibility, as age, school grade, degree of White blood?
4. Which of the above factors has the greatest influence on speed?
5. Does this study indicate that the Indian race is mentally retarded?

Materials

The Thorndike Handwriting Scale was used in measuring samples of handwriting of Whites and Indians. The rating was done by students in classes in educational measurements. A sample was often rated by as many as ten individuals, and the average of the ratings taken as the measure. The speed was determined by dividing the total number of letters by the total time of writing. Because the number of cases of the Whites was so small, we have been compelled to resort to the use of Starch's norms for school grades

COMPOSITION OF THE GROUPS AND THE RESULTS OF MEASUREMENT

Table I will show the racial and educational composition of the groups. There were six hundred three full-blood Indians and one hundred ninety-six mixed-blood Indians found in the United States Indian Schools at Chilocco, Oklahoma, and Albuquerque, New Mexico; and there were two hundred sixty White children taken from city and suburban schools of Colorado. Table II will show the medians of

TABLE I

Grade	4	5	6	7	8	Total
Full-bloods	37	64	129	101	182	603
Mixed-bloods	14	35	35	64	42	196
Whites.	38	58	61	68	45	260

TABLE II

Grade	4		5		6		7		8	
	Median	Quartile	Median	Quartile	Median	Quartile	Median	Quartile	Median	Quartile
Group:										
Full-bloods	14 8	2 2	15 8	1 7	15 0	1 6	15 5	1 5	16 6	1 2
Mixed-bloods .	16 2	1 2	14 7	1 3	16 8	1 1	16 1	1 2	16 7	8
Whites . .	10 1	7	10 9	8	12 1	6	12 6	5	13 6	8
Legibility.										
Full bloods	8 33	93	8 74	1 01	8 42	87	9 29	1 12	10 36	1 20
Mixed-bloods	8 33	96	7 81	89	8 75	84	8 81	95	9 91	1 10
Whites .	8 64	91	8 78	91	9 44	87	9 60	1 01	9 90	81
Starch's norms	8 7		9 3		9 8		10 5		10 9	
Full bloods SN overlap	41 per cent		20 per cent		14 per cent		21 per cent		37 per cent	
Mixed-bloods overlap	43 per cent		9 per cent		20 per cent		17 per cent		28 per cent	
Speed:										
Full-bloods	41 67	20 47	53 85	13 22	64 25	12 29	70 78	14 12	69 47	11 49
Mixed-bloods	45 00	12 21	47 72	9 75	57 20	11 40	72 5	11 34	64 29	10 50
Whites	33 08	8 8	46 92	10 57	46 74	11 17	54 29	15 52	68 11	15 39
Starch's norms	47		57		65		75		83	
Full-bloods SN overlap	46 per cent		44 per cent		48 per cent		44 per cent		19 per cent	
Mixed-bloods overlap	40 per cent		37 per cent		34 per cent		45 per cent		25 per cent	

ages and scores for the several grades along with the measures of overlapping, while Table III will show certain correlations found between the various factors under consideration, *i.e.*, legibility, speed, age, school grade, and degree of blood

Only when undertaking to find correlations between degree of blood and performance was the mixed-blood group enlarged by some Whites and full-blood Indians taken at random

TABLE III

	(1) Score	(2) Speed	(3) Age	(4) Grade	(5) Degree
Full bloods		r_{12} = 245 ± 027	r_{13} = 064 ± 027	r_{14} = 29 ± 021	
			r_{23} = 29 ± 03	r_{24} = 32 ± 024	
				r_{34} = 25 ± 025	
Mixed bloods		r_{12} = 199 ± 041	r_{13} = 027 ± 045	r_{14} = 42 ± 036	r_{15} = 051 ± 04
			r_{23} = 09 ± 044	r_{24} = 37 ± 038	r_{25} = - 05 ± 04
				r_{34} = 23 ± 042	r_{35} = 17 ± 04
					r_{45} = 008 ± 05
Whites		r_{12} = 205 ± 04	r_{13} = 294 ± 0.8	r_{14} = 45 ± 04	
			r_{23} = 44 ± 03	r_{24} = 65 ± 01	
				r_{34} = 92 ± 008	

TABLE IV. AS MULTIPLE CORRELATIONS FOR MIXED-BLOODS, WITH AGE ELIMINATED

I. Score legibility (1), grade (2), degree (3)

$R_1(23) = .41$ Using $x_1 = b_{12.3}x_2 + b_{13.2}x_3$

$x_1 = .38x_2 + .036x_3$

Weight of school grade = .38 Weight of degree of white blood = .036

II Score speed (1), grade (2), degree (3) $R_1(23) = .20$

$x_1 = b_{12.3}x_2 + b_{13.2}x_3$ $x_1 = .352x_2 - .057x_3$

Weight of school grade = .352. Weight of degree of White blood = -.057.

INTERPRETATION

It will be seen upon examination of Table II that the Indians, both full- and mixed-bloods, are much older than the Whites. For a school grade the ages of the full-bloods run on the average three years, and the mixed-bloods run four years older than the Whites. The range of age for full-bloods is from ten to twenty-three years, for mixed-bloods is from ten to twenty-three, and for the Whites is from eight to fifteen However, the median scores of legibility for the racial groups are just about the same, *i.e.*, full-bloods, 9.2; mixed-bloods, 8.8; and Whites 9.1. On the other hand the speed of the full-bloods is greater than that of either the mixed-bloods or the Whites. In turn the speed of the mixed-bloods is greater than that of the Whites these median scores for speed are respectively 60.9, 57.9, and 50.2 letters per minute. The reader must not conclude that the difference in speed is due to race as appears on the surface. This will be discussed later on.

It will be better to examine Table II for a comparison of legibility scores and speed scores by grades for the racial groups. It will be seen here that performances of the full-bloods are on the whole more nearly like the White norms from Starch, (as is indicated by the overlapping) than are those of the mixed-bloods. However, they both appear as rather inferior in performance to Starch's norms. They are more nearly like our medians for Whites. That is to say our Whites suffer in comparison with the Starch norms. If race made a difference, the mixed-bloods ought to show up better upon comparison than the full-bloods, but they do not. The same is true for speed. The full-blood performances are more nearly like the Starch norms for speed than are the mixed-bloods as judged by the overlapping.

If the reader will refer to Tables III and IV, he will see that the greatest correlation is found for all groups between legibility and school grade, and speed and school grade with exceptions found in the White data. Age is a small factor excepting in the case of the Whites where it is found to be significantly correlated with speed (.44) and slightly with legibility (.28). For Whites there is the usual high correlation between age and grade (.92). One should notice the insignificant correlation coefficients for degree of blood and Legibility and Speed. They are respectively .053 and .051. Multiple coefficients of correlation (Table IV) were found between legibility and school grade and degree, and speed, school grade and degree for the mixed-blood group. These are respectively .33 and .26. These are not very high. They were calculated for the purpose of determining the weight that should be assigned respectively to school grade and degree of White blood. For Legibility it will be seen that the ratio for school grade and degree is ten to one, and for speed, about six to one with the one a minus quantity. It will be seen that degree of blood is a matter of small significance. Elsewhere it has been shown that instead of degree of blood having the weight, it seems to have the situation may possibly be due to the influence of social status going with the degree of blood. At any rate degree is of extremely small weight.

SUMMARY

What, then, does this study show?

1. The full-bloods are on the whole somewhat better in legibility than the mixed-bloods, and the Whites are best of all.

2. The Indians are no more nor less speedy than Whites excepting in the Eighth Grade where the Whites excel.

3. The greatest factor in influencing score in legibility and speed for the Indians is school grade. Age has little significance, and degree of White blood practically no significance if any.

4. This study does not indicate that Indians are mentally retarded at least as this is shown in respect to ability in handwriting. It is seen that with proper training Indians may compare well with Whites in handwriting.

A MEASUREMENT OF THE KNOWLEDGE OF PSYCHOLOGY BEFORE AND AFTER FORMAL TRAINING

CALVIN HALL

University of California

The purpose of this investigation was five-fold. We wished to find: (1) The amount of improvement on a comprehensive objective examination in elementary psychology before and after taking a course in that subject; (2) the relation between the scores on the first and second taking of the test; (3) the relation between first and second taking of the test and the score on a college aptitude test; (4) the relation between gains—that is, score on the last examination minus score on the first examination—and college aptitude score; and (5) the effect of training on test reliability.

The first and last examinations were identical and were given on the initial and final meetings of the class. The examination consisted of one hundred eighteen true-false items based upon Perrin and Klein's "Psychology," the chief textbook used. The Washington College Aptitude test was given as a measure of college aptitude. This test is made up of the usual "intelligence" test items. The subjects were all members of a class in Psychology 1A at the University of California during the summer of 1940. This course extends over a period of six weeks.

The data are presented in Tables I and II.

TABLE I.—SUMMARY OF CORRELATIONS, MEANS, SIGMAS, STANDARD ERRORS OF SIGMAS, AND COEFFICIENTS OF VARIATION

Variable	N	Mean	SD	SD of dist	Coefficient of variation
College aptitude	108	124.62	28.328		
First examination	94	72.50	5.868	±.428	8.087
Last examination	94	92.76	6.928	±.505	7.469

The lack of or negligible correlations indicate that little relationship exists between the variables we were investigating. Considering the corrected coefficients alone, however, certain trends are noticeable. Thus, whatever is measured by the Washington College Aptitude test has more in common with the first presentation of the psychology examination than with the last. This is to be expected. "Intelligence" should be a more important determinant of information before than after formal training because other factors enter into the final examination, *e.g.*, interest, concentration, habits of study, to lessen the effect of "intelligence." The lack of correlation between "gains" and college aptitude score can be interpreted in two ways. Either improvement in this specific course is not dependent upon any of the factors that determine "intelligence" test score or the law of diminishing returns may operate to enable those doing poorly on the first taking of the examination to make greater gains than their more informed classmates The first and last taking of the examination correlate just moderately but probably as well as any two course examinations Certainly the score on our examination before formal training cannot be used to predict what score will be made on the same examination after training

TABLE II SUMMARY OF UNCORRECTED AND CORRECTED CORRELATION COEFFICIENTS AND RELIABILITY COEFFICIENTS WITH PROBABLE ERRORS (UNCORRECTED *r*'s IN PLAIN PRINT, CORRECTED *r*'s BY UNDERSCORING AND RELIABILITY *r*'s IN ITALIC

Variable	College aptitude	First examination	Last examination	Gains
College aptitude	*.875* ± *.019*	**.203**	**.162**	
First examination	.184 ± .068	*.400* ± *.080*	**.484**	
Last examination	.128 ± .069	.201 ± .065	*.790* ± *.042*	
Gains	.024 ± .069			*.665* ± *.044*

The objection may be raised that we did not validate our examination. This objection does not vitally affect our results, however We assumed that we were measuring knowledge of psychology when we selected questions from the textbook just as every instructor does when he constructs an achievement test for grading purposes.

Further conclusions that should be noted are the following

1. Although the absolute variability on the last examination increased over that of the first, the difference between the standard deviations is not significant. The relative variability as measured by

the coefficient of variation reveals the opposite tendency. The last examination was ninety-two per cent as variable as the first examination. This difference however is probably not significant.

2. The reliability coefficient of the last examination increased markedly over that of the first examination. This demonstrates the fact that the subjects tested as well as the inherent nature of the test may contribute to the size of the reliability coefficient. Hence any statement of a reliability coefficient for a given test should be accompanied by a description of the subjects.

3. The use of tests for predicting success in specific school courses is of doubtful value if the test is not standardized and possesses a low or unknown reliability. Such tests are often used as pedagogical devices for selecting students where the class is restricted and the writer on the basis of his data condemns this practice.